FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Improving Courses Management by Predicting the Number of Students

Vasco Taveira Gomes

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: João Mendes Moreira

Second Supervisor: João Pascoal Faria

June 24, 2016

Improving Courses Management by Predicting the Number of Students

Vasco Taveira Gomes

Mestrado Integrado em Engenharia Informática e Computação

June 24, 2016

Abstract

Every year, in higher education institutions all around the world, millions of students are required to choose the curricular units they are interested in enrolling for the coming semesters. When managing courses and their respective units, colleges and universities aim to predict and understand these demands in order to better plan the next scholar year. By successfully predicting students' demands, universities are able to ensure their limited budgets and resources are properly allocated. This study intends to answer the needs of a course administrator regarding student registrations by analyzing and applying predictive models. The main focus of this work was the prediction of the number of students enrolling in optional units, number of students per optional unit and number of students per non-optional unit in the syllabus. While some conclusions may be reached by simply measuring and extrapolating the difference in the number of students per year, identifying the cause of this difference is fundamental in the construction of a complete predictive model. Factors such as student grade average per unit or perceived difficulty, that are not always visible in the data, were also taken into consideration. For the development phase of the investigation, this project was applied to the course of Master in Informatics and Computing Engineering at the Faculty of Engineering of the University of Porto in the form of a case study. Multiple predictive algorithms, such as MARS, random forests and neural networks, were applied to each prediction topic. Using k-fold cross-validation, the models constructed were compared among themselves and against naive estimates based only on the number of students in previous occurrences. Overall, the best fit selected and presented for each topic was proven to surpass the naive alternatives.

Resumo

Todos os anos, em estabelecimentos de ensino superior de todo o mundo, milhões de estudantes participam no processo de seleção das unidade curriculares em que se irão inscrever durante os próximos semestres. Para uma gestão dos cursos e respetivas unidades, as universidades visam prever e compreender esta procura de modo a obter um melhor planeamento para o próximo ano letivo. Através de uma previsão bem sucedida da procura do corpo estudantil, as universidades são capazes de assegurar que os seus orçamentos e recursos são alocados corretamente. Este estudo pretende responder às necessidades de um administrador de curso relativamente ao registo de estudantes através da análise e aplicação de modelos preditivos. O foco principal deste trabalho foi a previsão do número de alunos a matricularem-se em disciplinas optativas, número de estudantes por disciplina optativa e número de estudantes por disciplina não optativa no currículo. Embora algumas conclusões possam ser alcançadas através da simples medição e extrapolação do número de estudantes por ano, a identificação da causa desta distinção é fulcral para a construção de um modelo preditivo completo. Fatores como a média académica por unidade ou dificuldade subjetiva, que nem sempre se encontram visíveis nos dados, devem, também, ser considerados. Para o fase de desenvolvimento da investigação, este projeto foi aplicado ao curso de Mestrado Integrado em Engenharia Informática e Computação da Faculdade de Engenharia da Universidade do Porto sob a forma de um caso de estudo. Vários algoritmos de previsão, tais como MARS, florestas aleatórias e redes neuronais, foram aplicados a cada tópico. Tendo por base validação cruzada, os modelos construídos foram comparados entre si e contra estimativas ingénuas baseadas apenas no número de estudantes em ocorrências anteriores. No final, o melhor modelo selecionado e apresentado para cada tópico provou superar as alternativas ingénuas.

Acknowledgements

I would like to preface this by expressing my sincere gratitude and appreciation to my supervisor, João Mendes Moreira, for the patience at guiding me along this subject and dissertation. It is still difficult to believe how far this opportunity has taken me. His expertise and invaluable advice were vital throughout this project.

I would also like to acknowledge João Pascoal Faria, my co-supervisor, and Susana Santos Gaio, from GSI, for their support with the case study. I would not have been able to complete this investigation without their continuous assistance.

To the friends who were always by my side, for their encouragement and motivation. To Bernardo, Daniel, Hugo, Ricardo, Nuno, for making this academic adventure so much more enjoyable. I genuinely hope your friendship accompanies me much further in the future. A special word goes out to Pedro Rodrigues, who kept pushing me to succeed in the most overwhelming of times.

Lastly, I would like to thank my family for their unconditional love and support over the past five years. To my uncle, who always believed in me. To my grandmother, the most generous person in the entire world. To my mother, for all she has given me, and for making me the person I am today.

Vasco Gomes

"Prediction is very difficult, especially about the future."

Niels Bohr, Nobel laureate in Physics

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

caret	Classification And Regression Training
CART	Classification and Regression Tree
CU	Curricular Unit
EDM	Educational Data Mining
FEUP	Faculty of Engineering of the University of Porto
GLM	Generalized Linear Model
GSI	Information Systems Office
HEI	Higher Education Institutions
k-NN	k-Nearest Neighbors
KDD	Knowledge Discovery in Databases
LA	Learning Analytics
MIEIC	Master in Informatics and Computing Engineering
MARS	Multivariate Adaptive Regression Splines
OLS	Ordinary Least Squares
PLS	Partial Least Squares
RMSE	Root Mean Squared Error
SVM	Support Vector Machine
UP	University of Porto

Chapter 1

Introduction

The current section serves as an introduction to the context in which this dissertation is inserted in regards to research area (Section 1.1), motivations (Section 1.2) and objectives (Section 1.3). A brief description on the case study used to validate the investigation is also provided. Finally, the document's structure and composition is presented (Section 1.4).

1.1 Context and Scope Overview

Higher Education Institutions (HEI) have provided an optional post-secondary stage of formal learning for over 50 years. These institutions, represented by academies, colleges, institutes of technology and universities, constitute the central establishments for the offering of higher and tertiary education around the world. HEI have seen a gradual increase in numbers over the past decades, with over 23 000 independent, accredited institutions as of 2015 [Lab15]. As highly complex organizational bodies, HEI are often structured as hierarchical entities, divided by organic units such as faculties or departments. Each unit is responsible for administrative decisions within its own scope, although general policies and processes should typically coincide with the approaches developed at upper levels.

In order to adequately measure and scale their growth, educational bodies have adopted the use of data exploration and analysis. This data, known as educational data, may encompass information with differing granularity collected from multiple hierarchical settings. Fine granularity implies comprehensive, detailed data such as student grade average and attendance per curricular unit. In contrast, coarse granularity represents comparatively simpler data such as student grade average per curricular semester. Exploring the underlying context and relationships embedded in this information allows institution administrators to make knowledgeable decisions about how to coordinate and manage their resources.

Educational data analysis is not, by itself, a recent topic, with early studies and hypothesis dating as far back as 1970 [Tin75]. However, the application of data mining and statistics tools

and methods is only now emerging in a research area known as Educational Data Mining (EDM). Recent studies have mostly focused on the topics of student attrition and retention, while some efforts have been placed towards the discovery of causes for academic performance [IR07].

Nowadays, the evolution of computer technology has cleared the path for new possibilities in data collection. Given the large amount of educational data and statistics available in most institutions, it is now possible to explore patterns in the data in order to gain insight into future occurrences. By successfully predicting future events, HEI are able to increase their administrative effectiveness. It can be argued that, as the amount of educational data increases, so does the potential for further analysis.

1.2 Motivation

Every year, all around the world, millions of students are required to choose the curricular units they are interested in enrolling for the coming semesters. When managing courses and their respective units, or subjects, HEI aim to predict and understand these demands in order to better plan the next scholar year. It is based on these predictions that institutions decide on how many registrations per curricular unit should be opened and how schedules should be distributed. By successfully predicting students' demands, institutions are able to ensure their limited budgets and resources are properly allocated.

This type of prediction can generally occur at multiples levels in an organizational hierarchy. Figure 1.1 illustrates a typical hierarchy of an institution. At the upper levels of the organization, an institution is managed by a rector, or chancellor, frequently accompanied by an executive board. Decisions, policies and processes defined at this level have an impact on the academy as a whole. The units below, usually structured as faculties, represent semi-independent bodies responsible for their own curricula. Faculties are presided by a dean and, in some cases, an auxiliary board of faculty professors. Each faculty may then comprise several courses from a broader scientific area, normally governed by a regent or course director. Some institutions also take into consideration department units, responsibility of heads or directors of departments. Departments are either structured as separate entities that co-exist along with courses, or placed at an upper hierarchy level and, therefore, directly responsible for them.

Lower hierarchical levels have a lower degree of internal variance with respects to their educational environment, and are associated with a narrower decision and responsibility scope. As depicted in Figure 1.1, a course represents the lowest unit in the educational chain, with each course director or administrator having the possibility of implementing individual policies for the management of curricular units and student registrations. It is, thus, appropriate to regard the prediction of student enrollments from the viewpoint of a course administrator. Higher administrative roles may then use these calculations as a baseline for larger scale predictions. It should be noted, however, that course-based predictions should later coincide with those formulated at an upper level. For instance, the sum of predictions of the number of enrolling students for all courses should be equal to the prediction of the total number of students enrolling in the faculty.

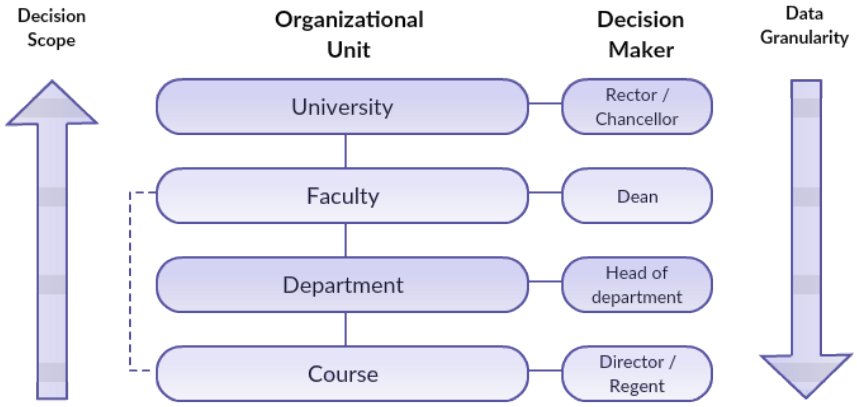

Figure 1.1: Hierarchy of a higher education institution

For the most part, the process for registration within curricular units covers three phases: students' selections, registration adjustment (sometimes referred to as *add and drop* period) and final placement. During the first phase, students are required to select the units they would like to attend during the coming semester. This selection is predominantly based on personal interest from the possible attendees after reviewing the syllabus for the subjects available. Due to constraints related to the number of possible registrations per unit and conflicting schedules, not all students will be allocated to the subjects they previously selected. When dealing with a number of applications that surpasses a subject's registration limit, faculties or individual courses may employ policies such as time of selection or student grade average to decide on which applications should take priority.

Registrations are later balanced during the adjustment phase, in which students that were not placed in one or more subjects are instructed to select new curricular units. Some faculties may also allow students to cancel their accepted registrations (an equivalent term for this process is *dropping* a subject) or replace them with new ones. The adjustment period may include multiple selection stages, until all students are enrolled in a minimum number of units.

The final placement phase constitutes the conclusion of the automated process. Afterwards, further placement changes are reviewed on a case-by-case basis by the course's administrative staff. It is common for this phase to last one or two weeks in the beginning of the semester, after which registrations are considered to be closed for the remainder of the school period.

Typically, the process of predicting the distribution of students per curricular units occurs before the semester begins, as noted in Figure 1.2. This timing emerges from the necessity of knowing, in advance, how the institution's resources should be allocated. Without this prediction, it would not be possible to ensure administrative requisites are fulfilled. These requisites may include, for instance, hiring the correct number of teachers or reserving the correct classes, as some subjects may require laboratories or rooms with specific technology. Furthermore, the prediction process needs to account for the time required to generate academic schedules. In larger faculties, the computation time for such a task may encompass several days. As the figure demonstrates, predictions for the second semester may be calculated once the information from the final placements

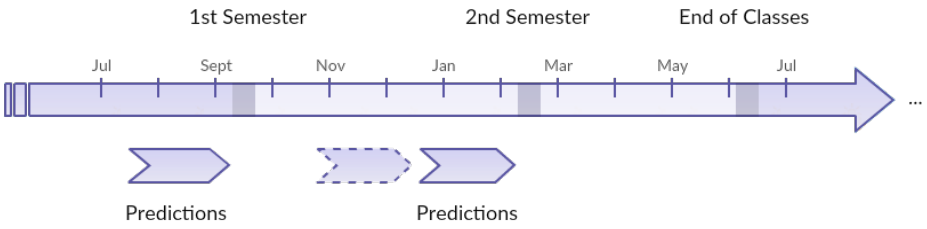

Figure 1.2: Timeline for predictions

of the previous semester is available, although this normally is not the case.

Nowadays, course administrators generally depend upon one of three simple approaches for the prediction of the number of students attending a curricular unit in a given semester: template and history-based solutions, student voting, or a combination of the two previous methods. Template and history-based solutions rely on data from previous semesters to derive basic formulas for the predictions. For instance, a 10% increase in registrations for a given subject in the two previous semesters could be an indication of a similar increase for the next one. Predictions can be inferred from the mean or trend in a variable number of templates. Student voting is an alternative that requires students to vote for the curricular units they are interested in before registrations take place. These approaches are not exclusive, and may be combined.

In most cases, template-based solutions are incapable of finding correlations unless explicitly implemented to do so. In these methods, a sudden reduction in registrations caused by a new teacher or a change in grading policy would not be able to be explained; in fact, those factors would not be inputs in the prediction. More accurately, template methods are fallible due to their lack of interpretability and adaptability to results that skew expected changes. Similarly, the reliance upon previous instances on a subject by subject basis represents a challenge when attempting to fit recent or emergent curricular units.

Student voting does not require a similar level of interpretability due to the inherent accuracy of the prediction as, in theory, votes should coincide with the registrations submitted for the final placement phase. Nonetheless, this is not the case in practice. Kardan et al. [KSGS13] propose several reasons for this incongruity, to note: (1) the students' lack of consideration for pre-requisites when voting, which later results in the impossibility of registration; (2) overlapping or conflicting schedules, which lead to the selection of a combination of curricular units that cannot occur in practice; (3) the students' lack of an unambiguous understanding of goals and priorities; (4) the lack of participation in the voting process, even when it is required. While these characteristics can be compensated by the integration with a template-based solution to a certain degree, an optimal solution should always account for all factors with the potential to affect the outcome.

In order to overcome these limitations and implement more comprehensive solutions, course administrators have started to resort to predictive analytics. Predictive models leverage statistics and machine learning techniques to make predictions about future behaviors and events. As a re-

sult of their flexibility, they can be applied to various industries. Predictive models are particularly sensitive to patterns in the data, being capable of perceiving correlations among multiple input factors. These models may also be interpretable, depending on the algorithm employed. Identifying and understanding which factors have the most influence towards the outcome is crucial, as it enables course administrators to assess what measures should be exercised so as to increase or reduce the number of registrations in a specific curricular unit. Factors may include course characteristics, instructor characteristics, subject workload or subject grade average [KSGS13]. Likewise, identifying further data that produces an influence in those factors and capturing it may prove helpful for the quality of the prediction [DPV09].

Predicting the number of students attending a curricular unit in a given semester is a critical step for the successful management of a course in a higher education institution. Failing to predict students' demands may lead to an inefficient allocation of the institution's resources, student dissatisfaction and an inherent decrease in grading results. Therefore, being able to identify and successfully predict students' interests is essential in order to achieve optimal results for both the management body and the students. In some sense, these predictions may also have an impact in an industrialized country's supply-chain. Elaborating plans to direct more students towards fields where professionals are needed may speed up technological research and development, enhancing a country's ability to compete in international economic markets [LLH12].

1.3 Proposal Definition and Objectives

This study intends to answer the needs of a course administrator regarding the number of students per semester and curricular unit by analyzing and applying multiple predictive models. Each requirement will be presented and treated separately, as an individual model with specific input factors.

Firstly, it is important to make a distinction between optional and non-optional curricular units, in which all students are obligated to enroll in. Consider a sample of 100 students enrolled in one optional unit and one non-optional unit, with a failure rate of 20% in both. The following semester, at least 20 students are expected to join the non-optional unit (disregarding students who drop out), as its completion is a graduation requirement. However, there is no lower limit for the optional unit, as students can simply select another subject. In reality, as students often resort to peer recommendations when selecting subjects, new students may even avoid that particular unit. As such, optional and non-optional curricular units should be treated by different predictive functions.

In the interest of finding the function which best fits the problem, several predictive models will be built and applied to each administrative requirement. Modeling performance will be estimated with cross-validation methods. During a later phase, models will be evaluated and compared with each other using criteria such as Root Mean Squared Error (RMSE).

1.3.1 Prediction Topics

For this investigation, predictions will be divided and focused into three separate topics: (1) number of students per non-optional curricular unit, (2) number of students enrolling in optional curricular units, and (3) number of students per optional curricular unit. All predictions will be applied to a single academic semester.

1.3.2 Application

For the development phase of the investigation, the predictive models and hypothesis here developed will be applied to the course of Master in Informatics and Computing Engineering at the Faculty of Engineering of the University of Porto in the form of a case study. The main dataset used in this paper was provided by FEUP's Informatics Center (identified by the Portuguese acronym of CICA). The dataset contains information from enrollments and student questionnaires relating to the MIEIC course from 2009 to 2015.

1.4 Document Structure

The remainder of this document is divided into six chapters. Chapter 2 presents existing literature and investigation in the area of educational data mining. Predictive analytics and its areas of application are identified, and some of the most common regression methods are exposed and compared. Additionally, key publications in the topic of prediction of student populations are reviewed. Chapter 3 describes the methodology used in the case study, and presents a detailed analysis of the datasets utilized. In Chapters 4, 5 and 6 the results obtained for the models developed in the case study are demonstrated and discussed, with each chapter pertaining to a distinct prediction topic. Chapter 7 concludes the study with an overview of the results of the dissertation as a whole and a discussion on suggestions and ideas for future work.

Chapter 2

Background and State of the Art

This chapter presents a review of the background and state of the art on the topics of machine learning and predictive analytics applied to educational systems. Section 2.1 introduces the concept of Educational Data Mining along with a brief description of its history and recent developments. Section 2.2 describes predictive analytics and regression, and follows with a summary of common predictive models and their characteristics. Research and applications of predictive analytics are detailed in Section 2.3.

2.1 Educational Data Mining

Over the last decades, the advent of computer technology has paved the way for an ever-growing expansion in data collection and exploration. The constant increase in data complexity and size eventually lead to the necessity of new, automated tools for assisting the process of information management. The term data mining was developed to refer to the process of discovering patterns [HTF09] and extracting or mining knowledge from large amounts of data [WF05]. While it is often a synonym with Knowledge Discovery in Databases (KDD), some authors have distinguished data mining as the step in the KDD process that consists of applying data analysis and discovery algorithms with the purpose of pattern extraction [FPSS96]. Although its definition and scope often vary, data mining can generally be classified as the sub-field of computer science that utilizes techniques of artificial intelligence, machine learning and statistics to perform automated extraction of knowledge from large amounts of data.

Recently, data mining practices have been applied to various business areas, with the transformation of raw data into interpretable information being shown to produce valuable results for decision and management support systems. Campbell and Oblinger [CDO07] initially defined academic analytics as the use of statistical and data mining techniques focused on helping faculty and advisors address student success. Further studies on the topic of data mining applications for education institutions contributed to a new area of research known as Educational Data Mining

(EDM). Romero and Ventura [RV10] define EDM as a research area that deals with the development of methods to analyze and explore data originating in an educational context. In other words, it is described as the application of data mining techniques to educational data.

EDM can be applied to multiple levels of an educational organization's hierarchy. Usually, it serves the purpose of evaluating and supporting the management of an academic programme or estimating the effectiveness of pedagogical strategies on student performance [RV07]. Romero and Ventura [RV07] propose the application of EDM oriented towards three different actors: (1) towards students, and aimed at providing instructors with recommendations on how to improve the learning process; (2) towards educators, and focused on course content, structure and learning effectiveness; (3) towards academics responsible and administrators, and aimed at optimizing institutional resources and educational programmes from a macro perspective. A second viewpoint is presented by Baker and Yacer [BY09], who categorize four areas of application: (1) improvement of student models; (2) discovery or improvement of models of a domain's knowledge structure; (3) study of pedagogic and educational support; (4) refinement and advancement of educational theory.

The application of data mining methodologies to educational data has also been referred to as Learning Analytics (LA). Similarly, LA is focused on the collection and analysis of data with the purpose of supporting and optimizing educational environments.

Although adopting considerably overlapping activities, EDM and LA have developed as separate research communities with distinct roots and approaches. Siemens and Baker [SB12] identify five key areas of difference between the two, including a considerably greater focus on automated paradigms of data discovery and a stronger emphasis on the analysis of individual components and relationships in EDM. Nonetheless, it is argued that, rather than strictly exclusive definitions, these distinctions represent broad trends in the communities. Essentially, both EDM and LA reflect an emerging necessity to develop methods capable of extracting information from and providing insight into data arising from educational settings.

Throughout the last several years, there have surfaced new approaches and procedures for the application of data mining technologies to educational fields. One of the most recent definitions for major categories of EDM methodologies is presented by Baker and Inventado [LW14], who discuss four key classes of methods: (1) prediction models; (2) structure discovery; (3) relationship mining; (4) discovery with models. Figure 2.1 functions as a summary of the subset of methods comprised in each category. In this context, prediction models employ predictive analytics to infer or predict one factor in the data from a combination of other variables. The topic is described below in further detail.

2.2 Predictive Analytics

Predictive analytics is the practice of extracting information from historical data using statistical, machine learning and data mining techniques to identify the likelihood of future or unknown events. The focus of predictive analytics is the study of relationships between explanatory and

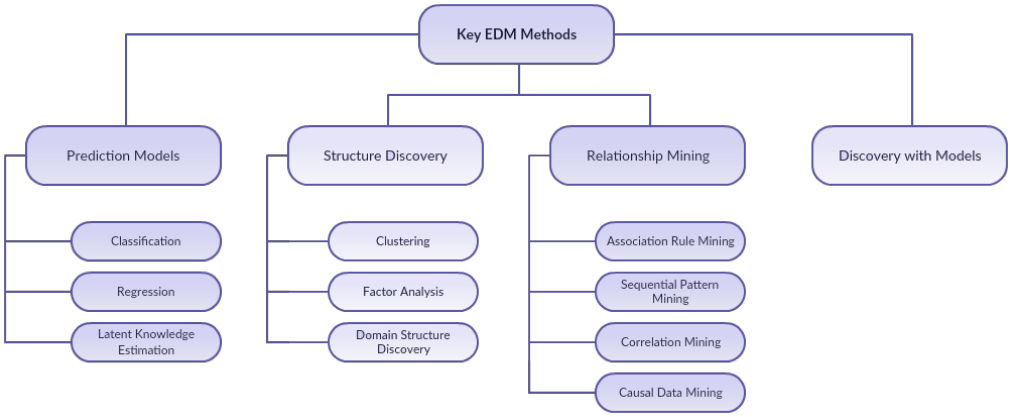

Figure 2.1: Key EDM methods (Baker and Inventado, 2014)

explained variables so as to assess or predict an unknown event with an accept level of reliability. Explanatory variables, also known as independent variables, represent potential causes for variation that can be measured individually. By combining these variables, one is able to explain or predict the outcome under study, known as the explained or dependent variable. Generally, predictive analytics encompass three different modeling areas: predictive modeling, descriptive modeling and decision modeling. It should be noted that, in some cases, a single model may be used to explain multiple dependent variables.

Descriptive models are centered on the quantification and segmentation of the data under analysis. These models are commonly utilized for data reduction, condensing big data into categorizable information. In descriptive modeling, the goal is to analyze past data to clarify what has happened and why it has happened.

Decision models are typically used in combination with business analytics to develop a formal model of a decision-making process based on multiple independent variables. These models can be used to provide insight into a decision and determine a course of action while maximizing specific outcomes. Recently, decision models have been used in conjunction with prescriptive analytics to support and take advantage of decisions based on predictive analytics.

Predictive models attempt to model the relationship between independent and dependent variables in order to reliably predict the probability of a given outcome. In most cases, predictive modeling is based on statistical methods supported by machine learning algorithms. A single predictive model may employ multiple algorithms and methods of statistical analysis.

The collection of historical data from which a model establishes relationships between variables is known as a dataset. The dataset lists the attributes, including dependent and independent variables, for a variable number of data samples, where each sample represents an individual observation in the data. For instance, in a dataset consisting of student characteristics such as yearly household income, age and academic average, a data sample would correspond to information about an individual student. The same information for all students would constitute the entire dataset. The process of assimilating knowledge from the dataset is known as the training phase.

Once the model has been built, the dataset may be used to assess and estimate prediction performance using validation techniques such as cross-validation.

Predictive models can be further divided by their level or interpretability in white and black-box models. White or glass-box models are interpretable in the sense that it is possible to examine the process that lead to a prediction. In black-box models, knowledge about internal mechanics is kept hidden, so interpretation is not possible (with an exception regarding model inputs such as independent variables). Due to this difference, white-box models are preferred when it is desirable to gain a qualitative understanding between independent variables and the outcome. It is important to mention, however, that a model's accuracy and performance is generally unrelated to its interpretability.

2.2.1 Regression Analysis

Regression approaches are some of the most common methods of predictive analytics. Traditionally, regression specifically refers to the prediction of a quantitative dependent variable. Accordingly, classification refers to the prediction of a qualitative response variable. Regression analysis is a statistical method for the investigation of relationships among variables. Regression may also be referred to as the study of how a quantitative response variable varies when an explanatory variable changes. In its simplest form, it attempts to find a function, referred to as the regression function, that fits a series of observations with minimal error. It should be noted that regression does not necessarily imply a causal relation between dependent and independent variables, but a significant association.

In these models, independent variables are often referred to as predictors. Analysis that assumes one independent variable is dependent upon one predictor is known as simple regression; analysis that assumes one independent variable is dependent upon more than one predictor is known as multiple regression.

Modeling regression can be approached with various methods. Traditionally, regression techniques are categorized as linear or nonlinear. In linear regression, it is assumed that the dependent variable is given by a linear combination of the parameters where each term is multiplied by a constant and then added to the result. In order to account for random noise that cannot be explained by the linear relationship, a variable known as model error or disturbance term is introduced. Linear regression models assume the form of the following expression:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_i x_i + e \tag{2.1}$$

where y represents the numeric response for the dependent variable, b_0 represents the estimation for a constant intercept, x_i represents the value of the ith predictor, b_i represents the coefficient for the ith predictor, and e represents the disturbance term. When a model can be written in this form, it is said to be linear in the parameters. Linear regression models attempt to

determine these parameters by minimizing the function in regards to an expression that compares observed with predicted values (the least squares method is a typical approach for this).

Generally, when using linear regression models, it is simple to interpret the relationship between dependent variable and predictors and analyze the correlation among predictors [KJ13]. Common regression models of this type include Ordinary Least Squares (OLS) and Partial Least Squares (PLS) regression.

Nonlinear regression models are expressed by functions that are not linear in the parameters. These models tend to vary in approach, and comprise many different techniques. Common models of this type include neural networks, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Multivariate Adaptive Regression Splines (MARS) and tree-based models.

2.2.2 Summary of Common Regression Models

The selection of a regression model is one of the most important steps towards developing a successful prediction. No model is uniformly better than all the others [KJ13], so one should conduct a proper review of the problem's characteristics before attempting to apply any given algorithm.

Most regression methods have a set of characteristics that distinguish them from the rest in regards to how they handle individual steps of the predictive process. For instance, some regression models are incapable of dealing with data in which the number of predictors per sample is larger than the overall number of samples, while others are not robust to outliers. Therefore, some models are inherently better suited for some situations than others. However, it should be noted that, while these characteristics generally hold, they are not applicable in every problem. It is seldom known beforehand which method will perform best for any given problem [HTF09], so these should only serve as guide.

Table 2.1 (adapted from Kuhn and Johnson [KJ13] and Hastie et al [HTF09]) presents some characteristics of the most common regression models. It is shown that interpretability, computation time and robustness to noise vary greatly per model, even among those in the same category. The column referring to tree-based models comprises the general characteristics of this family of predictive models, and each individual model may present different results.

2.3 Applications of Predictive Analytics on Academic Environments

In 2007, Romero and Ventura published a survey on educational data mining from 1995 to 2005 [RV07]. In 2009, Baker and Yacef [BY09] categorized the methods used in the research papers analyzed by Romero and Venture as (1) relationship mining, (2) prediction, (3) clustering, (4) human judgment or (5) neither, noting that papers could cover multiple categories. Of the 60 research papers, 17 (28%) involved predictive methods, making prediction the second most prominent area and only surpassed by relationship mining approaches (43%). Over the years, a new pattern emerged, with prediction gaining more focus. Baker and Yacef noted that predictive approaches became the most prominent area in the proceedings of the Educational Data Mining

Table 2.1: Summary of common regression models and their characteristics

	Predictive Model					
Characteristic	Neural Net	SVM	Trees	MARS	k-NN	L. Regression
Interpretability	✗	✗	✦	✓	✗	✓
Predictive power	✓	✓	✗	✦	✓	✦
Computation time	✗	✗	✦	✦	✓	✓
Robustness to noise/outliers	✗	✗	✓	✦	✓	✗
Handling of multiple data types	✗	✗	✓	✓	✗	✗
Handling of missing values	✗	✗	✓	✓	✓	✗
Allows # samples $<$ # predictors	✓	✓	✓	✓	✓	✗
# Tuning parameters	2	1-3	0-3	1-2	1	0
Symbols: positive (✓), negative (✗) and reasonable or in between (✦)						

Conference in 2008 and 2009, being present in 42% of the papers. Baker and Yacef also stated that the proceedings of both conferences accounted for approximately as many papers as those analyzed by Romero and Ventura from 1995 to 2005.

Predictive efforts have mostly been directed towards the prediction of student performance and attrition at various hierarchical scopes. Student attrition can be defined as the gradual reduction in the number of students enrolled in an education institution. Attrition is often mentioned in conjunction with retention, which is more attentive of the means on how to reduce student attrition. Historically, both terms have been used when referring to studies that aim to perceive the causes for a decrease in graduation rates. Research has shown that identifying such causes may support and improve educational effectiveness, enabling universities to react proactively in response to students at risk [Lua02]. Some studies have attempted to find a correlation between student performance and attrition, attributing performance as a possible cause of the attrition phenomenon. Other causes may include educational and financial demographics [LG08] [Del11].

The application of a predictive modeling approach for the management and administration of student populations is still vastly unexplored, with most studies dating from the past few years. In particular, the analysis of the number of students per academic semester or curricular unit is a recent topic with many investigation paths available.

Rosa and Pereira [RP13] proposed a metapopulational model for the study of the evolution of the number of students in an education institution. The approach describes different dynamics for the representation of the flow of the number of students entering and leaving the institution, and transiting from academic year within the establishment. However, the proposal does not contemplate variation in regards to the distribution of students per courses or curricular units.

N. and K. Patanarapeelert [PP13] suggest an alternative approach to the modeling of student populations from an institution's department level by employing regression analysis. Two models are formulated, respectively categorized as descriptive and explanatory. As the proposal is presented with regards to a department's administrative scope, individual curricular units are not reviewed in the solution.

Kardan et al. [KSGS13] present a modeling approach to course selection in online higher education institutions using neural networks. The study proposes and compares several predictors, such as course characteristics and student's workload, to predict the number of students per semester and online programme (equivalent to a traditional course). Additionally, three other machine learning techniques (SVM, k-NN and decision trees) are briefly described and utilized to compare and estimate levels of performance across techniques.

As evidenced, investigation on the application of predictive analytics to the modeling of student populations in regards to the necessities of a course's administration is a topic that hasn't been thoroughly explored. Further research in the area could be applied to the selection and comparison of predictors in traditional educational environments (versus online) and analysis of multiple predictive methods. It can be argued that future studies in the topic could prove beneficial for both the management of a course and the effectiveness of the academic programme as a whole.

Chapter 3

Methodology and Case Study

This chapter serves to formally describe the case study and the methodology employed in the construction of predictive models. Section 3.1 presents the context in which the case study is inserted and introduces all prediction topics in detail. In Section 3.2, modeling approaches and validation techniques are discussed and their selection is explained. Section 3.3 illustrates the process utilized in the dataset's exploration and preparation. Lastly, a brief overview of the software and working environment used for the modeling phase is given in Section 3.4.

3.1 Faculty of Engineering of the University of Porto

Founded in 1911 at the city of Porto, Portugal, the University of Porto (UP) is one of the largest higher education institutions in the country. As of 2015, the university is composed of 13 faculties, a biomedical sciences institute and a business school, and is responsible for an average of 30,000 students a year. The university also functions as home and collaborative partner to multiple research and development centers and institutes.

The Faculty of Engineering of the University of Porto (FEUP) was established as an engineering faculty under UP in 1926. The faculty is organized by engineering fields, including chemical and electrical engineering, and structured in several individual departments.

One of the academic degrees awarded by FEUP is the Master in Informatics and Computing Engineering (MIEIC), which combines a Bachelor and Master's degrees for a duration of five years. Integrated in the Informatics Engineering department of the faculty, MIEIC was formally established in 2006. In the academic year of 2015/2016, there were over 700 students enrolled in the course.

In a course such as MIEIC, a regent or course director is directly responsible for the management of most administrative requirements. In order to plan a successful scholar year, measures must be taken so as to properly govern the allocation of the faculty's resources. It is, therefore, necessary to understand how to organize curricular units in the syllabus, how to optimally distribute

classes and instructors, and how to accommodate the requests of the student body. Generally, all these conditions must be satisfied before the beginning of the semester.

3.1.1 Prediction Topics

This dissertation presents an approach based on predictive analytics on how to support the administrative necessities of a course director. It is demonstrated that the application of methodologies based on data exploration in order to generate predictions may have a positive impact on the management effectiveness of a higher education institution. To validate the hypothesis here presented, the models developed are applied, in the form of a case study, to the MIEIC course at FEUP.

Three prediction topics are analyzed: (1) number of students per non-optional curricular unit, (2) number of students enrolling in optional curricular units, and (3) number of students per optional curricular unit. The topics are analyzed separately, and different models are formulated for each case.

3.1.1.1 Number of Students per Non-Optional Curricular Unit

This study is directed at supporting the process of decision on the number of classes and teaching hours to allocate for a given subject. Conventionally, education institutions define policies limiting the number of students per class for any given subject. These policies are typically implemented at faculties or universities, and must be followed by its respective courses. In some instances, legislation may mandate a maximum student-teacher ratio for a particular learning stage.

For this process, it is paramount to understand trends in the data such as an increase in the number of students per unit over the years. Similarly, sequences in the data are of equal importance in the prediction, as is evidenced when retention numbers are considered. For instance, if 20 students fail to obtain the minimum grade in a given subject, at minimum, those same 20 students are expected to register for the subject's next occurrence. As such, these conditions imply the need for a predictive model capable of handling sequential data.

For this topic, each data sample will contain information about a specific curricular unit in a given year or edition. The model variables for a sample are identified as follows:

- Curricular unit ID;
- Number of registrations;
- Number of students approved;
- Grade average;
- Grade standard deviation;
- Number of students registered in the cycle of studies.

The cycle of studies refers to the degree or programme in which a student is participating. All courses are integrated within a degree, such as a Bachelor or PhD. For the case study, this will represent the students registered in MIEIC during a specific semester.

3.1.1.2 Number of Students Enrolling in Optional Curricular Units

This analysis is focused on supporting the process of decision on the number of optional curricular units to allocate for a given semester. Due to staff and budget constraints, not all subjects in the syllabus can be made available at a time. Consequently, course administrators must evaluate how many subjects should be opened based on the number of students expected to register.

For this topic, each data sample will contain information on a semester from a given year or edition. The model variables for a sample are identified as follows:

- Number of applications in optional units;
- Number of students registered in the cycle of studies;
- Number of students from the given course participating in mobility programmes at other faculties;
- Number of students from other faculties participating in mobility programmes at the given course;
- Mean number of delayed units per student.

A delayed curricular unit represents a subject the student has failed to complete and must, thus, enroll in again. In non-optional units, students may select other subjects with an equivalent number of credits. For this study, only the number of units is considered, as the subjects themselves are interchangeable and, therefore, irrelevant for the prediction.

3.1.1.3 Number of Students per Optional Curricular Unit

This research is aimed at supporting the process of decision on which optional curricular units to allocate for a given semester. Once the number of units has been determined, course administrators must evaluate which subjects to open based on possible student interest. Following this process, curricular units predicted to have a higher number of applications are made available for student registration. Note that the objective of this topic is not the prediction of the number of student registrations, unlike the previous studies, but the number of student applications.

In addition, identifying the causes for such interest may prove of significance in the process of managing individual curricular units. Rather than only perceiving trends in the data, it is equally important to distinguish the factors relevant to a student's selection of a subject. These conditions suggest the need for an interpretable predictive model.

Generally, template-based solutions that simply measure and extrapolate the difference in the number of students per semester are incapable of assessing which factors have the most weight in

the final result. In the construction of a complete predictive model, variables such as grade average per unit, perceived difficulty and subject workload, that are not always visible in the data, should also be taken into consideration.

For this topic, each data sample will contain information about a specific curricular unit in a given year or edition. The model variables for a sample are identified as follows:

- Number of applications;
- Number of registrations;
- Grade average;
- Grade standard deviation;
- Instructor(s) identification (academic code);
- Mean and standard deviation of the following topics from student questionnaires:
  - Perceived difficulty level;
  - Global appreciation of the unit;
  - Global appreciation of the instructor(s).

3.2 Modeling Methodology

The prediction topics here presented will be approached individually, as independent regression problems with no direct correlations. Each topic is to have its own input data, pre-processing phase, and model selection according to its specific requirements. Conclusions will be presented per topic and, later, as a whole in regards to the impact of predictive analytics on the case study.

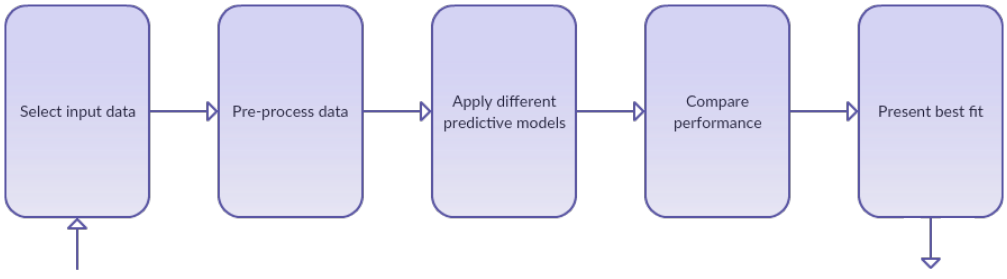

Figure 3.1: Modeling process applied to an individual prediction topic

The experimental process will be similar for all topics. Once the predictive models have been selected and trained, they will be compared against a predefined set of metrics. It should be noted that, in order to fully capture the potential of each model, the input data may be filtered or transformed by different methods for different models. Figure 3.1 denotes the general process that will be applied to all topics.

3.2.1 Model Performance Metrics

When utilizing predictive models, it is essential to identify which criteria to select for the measurement of performance and accuracy. For regression, where the response variable is a continuous numerical outcome, several estimations may be applied. The principal evaluation methods employed in this study are briefly described below:

- **Root Mean Squared Error (RMSE)**: measures the difference between values observed and values predicted by a model. The RMSE is one of the standard statistical metrics to measure model performance [CD14]. It is given in the form of a positive number in the same units as the original data, and may be interpreted as the average distance between observations and predictions. The RMSE is calculated with the square root of the mean of the squared residuals, where a residual represents the difference between an observation and corresponding prediction. The formula for calculating the RMSE is given by the following expression:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{3.1}$$

  where e represents the error, or residual, for a given observation, while n represents the total number of observations.

- **Coefficient of Determination ($\mathbf{R}^2$)**: summarizes the proportion of variance of the dependent variable that is explained by the regression model. R^2 is given in a range from 0 to 1, with 1 implying that the model can fully explain the variation in the outcome. It is important to clarify, however, that R^2 is a measurement of correlation, and not accuracy [KJ13]. Although it encompasses multiple equivalent definitions and formulas, the most common one is expressed below:

$$R^2 = \frac{SS_{res}}{SS_{tot}} \tag{3.2}$$

  where SS_{res} and SS_{tot} represent the regression sum of squares and the total sum of squares, respectively.

- **Naive Average**: functions as a baseline on which to test the model against. In this method, each prediction is assumed to be equal to the mean of the observations. A generic formula for its calculation is given by the following:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} y \tag{3.3}$$

where $\hat{y}$ represents the prediction value, y represents an observation, and n represents the total number of observations.

3.2.2 Model Validation

In order to reliably estimate a predictive model's performance, it is crucial to understand the conditions in which it will be tested. One of the fundamental requirements of a predictive model is its ability to generalize to previously unseen data. Put differently, a model should be able to capture the underlying relationships in the data that correlate to a change in the studied outcome, so that it can potentially fit new samples. For this reason, using the same data samples for the training and testing phases may result in biased evaluations and overly optimistic measurements of accuracy. Consequently, it is often necessary to split the original dataset in different partitions.

Generally, the dataset is partitioned in two or three different blocks: training set, testing set, and an optional validation set. The training set contains the data samples initially presented to the model so that it can reasonably estimate the parameters used in the construction of a prediction. In a linear regression model, for instance, the training set is used to estimate the coefficients for each individual predictor. Once the model has been trained, its performance can be assessed with the testing set. An optional partition, denominated validation set, may be utilized to tune a model's parameters or to simulate comparisons between multiple models before using the testing set. During the validation step, models are regularly adjusted with the objective of attaining more accurate predictions for samples present in the validation set. Once this stage is complete, the testing set serves to represent real-world data and confirm how the model would generalize when handling new data.

The process of splitting a dataset, by itself, may encompass a multitude of different methods and strategies. In some cases, a dataset may lack sufficient data such that the removal of samples from the training set may compromise the predictive ability of the model. For those cases, some authors [KJ13] advocate the use of resampling techniques over a single validation or test set. Most resampling techniques operate in the same manner: a subset of samples are used to train a model, while the remaining are utilized to estimate its performance. The process is repeated several times, and the estimations are then averaged and summarized. This functions as if the model had been validated against multiple, different test sets.

Due to the dimension of the dataset provided for the case study, a resampling method was selected over a traditional division in training and test sets. All the models described in this paper were evaluated using a variant of k-fold cross-validation. In k-fold cross-validation, the original data is randomly partitioned in k subsamples of equal size. $k-1$ subsamples are then used to train a model, while the remaining subsample serves as the validation set used to estimate performance. The cross-validation process is repeated k times, with each iteration representing a *fold*, and each subsample is used as the validation set once (and included in the training set $k-1$ times). Figure 3.2 depicts the process when $k = 5$.

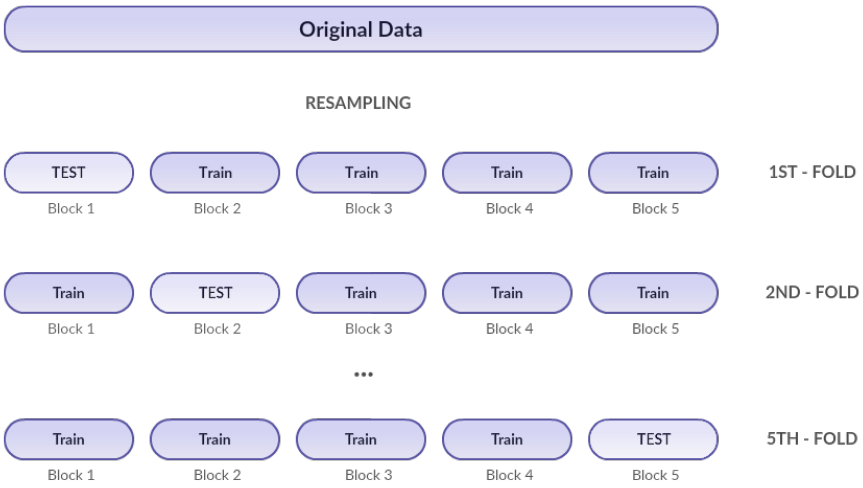

Figure 3.2: Schematic of the cross-validation method with 5 folds

The subsamples used for the cross-validation process are randomly generated, and the distribution of the response variable per fold may vary. To ensure reproducible results, a number seed is used for the random generator prior to the data splitting process.

Models are individually tuned according to the results obtained in the cross-validation process. Once the process is concluded, models are compared using the RMSE metric. The coefficient of determination for each model is also calculated and presented, but it is not considered in the comparison. Finally, the best models selected for each prediction topics are compared against a naive average. Figure 3.3 presents the approach described. Results for each topic are presented individually in their corresponding sections.

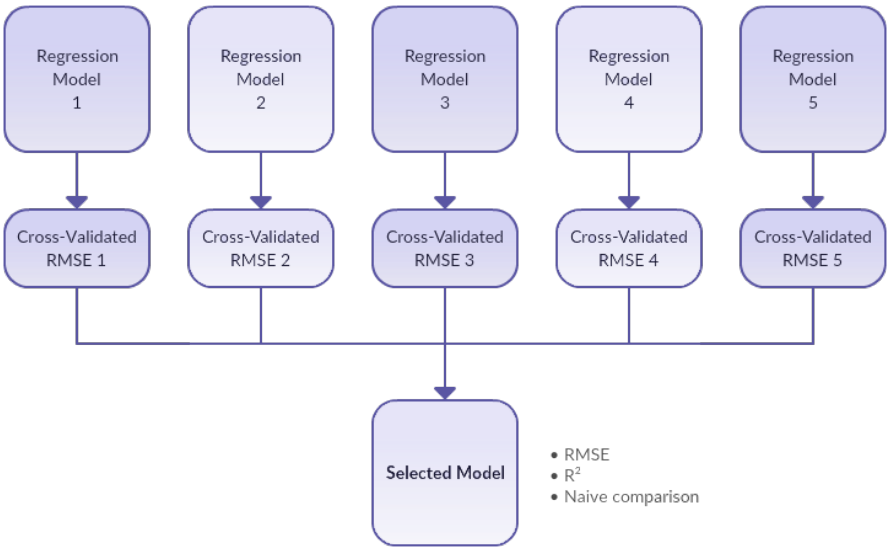

Figure 3.3: General model selection process

3.2.3 Predictive Models

Multiple regression models were tested and compared for each prediction topic. The major predictive models and algorithms utilized in this study are briefly described, individually, below.

3.2.3.1 Linear Regression and Derivatives

- **Linear Regression**: an easily interpretable model that attempts to minimize the sum of squared residuals. Also referred to as ordinary linear regression or Ordinary Least Squares (OLS). Ordinary linear regression is unable to find a unique set of regression coefficients when the number of predictors surpasses the number of observations.

- **Partial Least Squares**: in its simplest form, Partial Least Squares (PLS) seeks to find linear combinations of the original predictors, known as components, that maximally summarize predictor variability while simultaneously maximizing the components' correlation with the response variable. Much like OLS, it is considered an interpretable model.

- **Elastic Net**: produces biased parameter estimates by introducing penalties to the sum of squared residuals. Statistically, the increased bias may be able to reduce the mean squared error. The elastic net model is also referred to as elastic net regularization, as it attempts to control or regularize parameter estimates.

3.2.3.2 Nonlinear Regression Models

- **Multivariate Adaptive Regression Splines (MARS)**: partitions the original data into separate piecewise linear segments (splines). Given a set cut point, a predictor is separated into

two groups which are then used as features in a linear regression model. Predictor and cut point combinations are added based on the error of the generated model. Once all features have been selected, a pruning process may take place so as to remove features that do not contribute significantly to the final model.

- **Support Vector Machines (SVM)**: based on the concept of decision boundaries separating data points. To facilitate the process, inputs are mapped into high-dimensional feature spaces using a set of mathematical functions known as kernels. SVMs comprise a multitude of learning models and techniques, the one utilized in this study being known as ε*-insensitive regression*.

- **Neural Networks**: utilizes linear combinations of the original predictors, typically called hidden units, which are then transformed by a nonlinear function. The model's response is given a linear combination of the hidden units. A weight decay parameter, similar to the penalty discussed in the Elastic Net algorithm, is used to regularize over-fitting.

- **k-Nearest Neighbors**: uses a weighted average of the k closest data samples in the feature space. Distance between samples may be calculated by various metrics, such as the Euclidean or Mahalanobis distances.

3.2.3.3 Tree-Based Models

- **Regression Trees**: a tree-based model in which leaves represent possible responses and edges represent logical conditions known as splits. The number of possible predicted responses are finite, as predictions are calculated from the average of the samples in each terminal node. To avoid over-fitting, trees may be pruned using a complexity parameter derived from their depth.

- **Conditional Inference Trees**: an approach to the construction of basic regression trees which conducts statistical hypothesis tests for the selection of predictor splits. In this method, trees do not undergo a pruning phase.

- **Model Trees**: an extension to the basic regression tree in which a linear model is trained at every node. Predictions are given by a combination of the predictions from the models that define a path. The complexity of the final tree can be further reduced by removing specific conditions that do not contribute significantly to the model (rule-based approach).

- **Cubist**: a rule-based model developed as an augmentation to model trees (specifically, as an extension to Quinlan's M5 algorithm [KWKC12]). It utilizes a boosting-like scheme, called *committees*, where model trees are created in succession and adjusted as per their predictions on the training set. Cubist may also use nearby data points, the nearest neighbors, to tune predictions.

- **Bagged Trees**: utilizes multiple trees trained with distinct random samples of the original data. Samples are collected via statistical bootstrapping, a resampling technique with replacement in which the same observation may be included multiple times. The bagged model's prediction is given by the average of the combined predictions generated by each individual model. Bagging, an abbreviation for *bootstrap aggregation*, is an ensemble method, a technique which combines multiple models or learning algorithms.
- **Random Forests**: an extension to the bagging process which selects a random subset of the original predictors at each tree split in the learning phase. This addition reduces or eliminates possible correlations originated from the bagging process in which multiple trees acquire similar structures due to the selection of the same features.
- **Boosting**: combines, or boosts, weak learners such as trees to produce an ensemble. Using a loss function like the RMSE, boosting attempts to find an additive model that minimizes the loss function.

3.3 Data Analysis

Before experimenting with regression methods, a pre-processing step is required in order to properly format the data for the desired goals. The sources for this paper's data and preparation processes are described below.

3.3.1 Data Sources

The data used in this case study is a combination of two datasets provided by FEUP. The main source of data was provided by FEUP's Informatics Center (CICA), with the help of the Information Systems Office (identified by the Portuguese acronym of GSI). An additional, complementary dataset was supplied by MIEIC's course director. CICA's dataset is composed of multiple individual Microsoft Excel files, structured as follows:

- Students registered per curricular unit;
- Grade average and standard deviation per curricular unit;
- Instructor(s) identification per curricular unit;
- Students registered per semester;
- Students participating in mobility programmes per semester;
- Student questionnaires per curricular unit (2013/2014 to 2015/2016), with results for the following topics:
  - Autonomy support;
  - Consistency and help;

  - Structure;
  - Relationship;
  - Engagement;
  - Appreciation and clarity;
  - Evaluation;
  - Difficulty;
  - Impact.

- Applications per optional curricular unit (2014/2015 to 2015/2016);
- Mean number of delayed units per student (2015/2016).

Each of the above topics contains data from the academic years of 2009/2010 to 2015/2016 unless specifically stated otherwise. In its entirety, the dataset comprises twenty seven Excel Workbook files (*.XLSX*). Generally, a file represents the data for a single topic in a given academic year. In some cases, the workbooks contained several worksheets encompassing multiple years. The inconsistencies in the length of the data for each topic are due to a lack of information in CICA's system; the applications per optional curricular unit, for instance, were not available for previous years. Table 3.1 depicts an exemplar sample extracted from the file *inscritos2009*, which represents the students registered per curricular unit in the academic year of 2009/2010. All the data pertains to the MIEIC course.

Table 3.1: An exemplar sample from CICA's original dataset. The data represents the students registered per curricular unit in the academic year of 2009/2010.

	Data Columns				
Sample	Course	CU Code	Initials	Semester	Students
1	MIEIC	EIC0003	ALGE	1S	181
2	MIEIC	EIC0004	AMAT	1S	239
3	MIEIC	EIC0005	FPRO	1S	156
4	MIEIC	EIC0011	MDIS	1S	172
5	MIEIC	EIC0013	AEDA	1S	134
6	MIEIC	EIC0014	FISI2	1S	185
7	MIEIC	EIC0016	MPCP	1S	266
8	MIEIC	EIC0021	MNUM	1S	169
9	MIEIC	EIC0022	TCOM	1S	155
10	MIEIC	EIC0023	BDAD	1S	122

The dataset provided by the regent is structured in a single Excel Workbook. It contains information for every curricular unit (CU) from 2009/2010 to 2015/2016, with the following topics:

- Students registered;
- Students evaluated;
- Students approved;
- Grade average and standard deviation for evaluated students;
- Grade average and standard deviation for approved students.

The necessity for a distinction between registered and evaluated students emerges from the fact that some students may drop out of the unit (either officially, after requesting authorization from the course's administration, or by accumulating absences).

It should be noted that, in order to preserve a minimum level of anonymity, the instructors' identifications were supplied via their academic code. However, for the student questionnaires, this was not the case, and only names were provided. To circumvent this problem, an additional file with the association between names and academic codes for all the instructors in the faculty was provided by MIEIC's course director.

3.3.2 Data Preparation

Initially, most files were combined to facilitate the assimilation of the data into the working environment. Due to differences between formats, the process was mainly conducted manually. The files pertaining to the same topic were first merged and added to a single Excel worksheet, and all worksheets were later added to a workbook. To accelerate this process, the merging operations of some files were automated with scripts written in the R language.

Excel's internal functions such as *VLOOKUP*, capable of cross-referencing other worksheets via search/match criteria, enabled the construction of the two final datasets utilized in the study. The two datasets contain information for all curricular units and semesters from 2009/2010 to 2015/2016, respectively.

3.3.2.1 Curricular Units Dataset

The final dataset for curricular units is composed of a total of 463 individual entries. Of those, 29 entries represent units from the first semester of 2015/2016, and lack information in 6 columns (number of evaluated and approved students, and corresponding grade average and standard deviation). Table 3.2 presents the dataset's structure and an example of a random sample from the academic year of 2014/2015.

It should be noted that the column *course* was entirely removed, as it only introduces redundancy in the sense that all entries belong to the same course. Similarly, the column *initials* does not add any new information when combined with *code*, but it was maintained to simplify data exploration activities. The dataset also illustrates a distinction between year and curricular year: the column *year* denotes the current academic year, while *curricular.year* specifically refers to the

year in which a student in registered within the course. MIEIC has a duration of five years and has, thus, five curricular years. This is the terminology that will be used henceforth.

Additionally, it is important to explain that some incongruities were found in the number of students registered per curricular unit in the two original datasets provided for the case study. In roughly 25% of the entries, there was a difference between datasets of more than one registration in the number of students registered per curricular unit; of those, 15 entries had a difference of more than five registrations. In such cases, the dataset provided by the regent was given precedence, so as to maintain consistency with the number of evaluated and approved students - these numbers are not available in CICA's dataset. It is believed that these differences are due to the date at which the datasets' images were created, as this information is regularly updated. It is not clear, however, which dataset is most recent.

Lastly, there were two cases, relating to the subjects *Dissertation* and *Preparation of the Dissertation* that were entirely removed from the data. These curricular units occur twice per year, unlike regular units that only take place once per edition, which lead to multiple conflicts in the information available. As the datasets originally provided were often discordant in regards to the variables that characterize either unit, their occurrences were set aside.

Table 3.2: Structure for the curricular units dataset

Column	Example
year	2014
code	EIC0078
initials	TNEL
curricular.year	4
semester	2S
registered	25
evaluated	25
approved	25
evaluated.avg	16.08
evaluated.sd	1.06
approved.avg	16.08
approved.sd	1.06
registered.semester	634
opt	Y
candidates	25
teacher	211625

3.3.2.2 Semesters Dataset

The final dataset for semesters is composed of a total of 71 entries, with each sample referring to the data of a curricular year in a given semester. As an example, there are generally five entries for any given semester (corresponding to all curricular years in the syllabus). All entries are complete with values for all columns.

Table 3.3 presents the dataset's structure and an example of a random sample from the academic year of 2010/2011. It should be taken into account that, while the column *students* refers to students registered in the fifth curricular year, the columns for *mobility* include students registered in any year; this is not a pondered decision, but a limitation set by the information available.

Table 3.3: Structure for the semesters dataset

Column	Example
year	2010
curricular.year	5
semester	1S
students	569
mobility.in	13
mobility.out	14

The two datasets presented in this section (curricular units and semesters) are contained in a single Excel Workbook as two individual worksheets. No other data is used in the construction of the predictive models.

3.4 Tools and Software

The modeling phase of this study was entirely conducted in the R language. R is a programming language and software environment focused on providing a wide variety of facilities for statistical computing and graphics [VS11]. It integrates a diverse set of specialized techniques, including linear and nonlinear modeling, clustering or time-series analysis, and its capabilities are greatly extended with software packages. Packages are available as libraries mainly focused on specialized topics and techniques.

All the models here presented were built and trained via the Classification And Regression Training (*caret*) package. As per its author, *caret* contains a set of functions that attempt to streamline the process for creating predictive models in regards to data splitting, pre-processing, feature selection and model tuning. Internally, *caret* imports over a dozen other packages to fit specific requirements (*earth*, *nnet*, *party* and *Cubist*, among many others).

Although R provides direct support for Excel files, workbook files (*.XLSX*) are not included. As an alternative, the package *XLConnect* was used. The datasets are loaded directly into data frames, one of R's default structures and the one most indicated for storing data tables [VS11].

This ensures all data can be easily edited in Excel and loaded into R without requiring an intermediate conversion.

When referring to execution times in later sections, it should be noted that all models were constructed with an Intel Core i5-4210H processor and 8GB of RAM. Models were tested under both RGui and RStudio.

Chapter 4

Number of Students per Non-Optional Curricular Unit

This chapter presents the results obtained with the predictive models constructed for the prediction of the number of students registered per non-optional curricular unit. Section 4.1 describes, in detail, the data structure selected for the topic. Section 4.2 illustrates the process applied to and results obtained by each individual model. Further experiments and novel approaches are discussed in Section 4.3. Finally, Section 4.4 concludes on which model provides a better fit to the problem.

4.1 Experimental Setup

As previously explained, all students are required to complete the non-optional curricular units in the syllabus. It was evidenced that sequences in the data are crucial to fully represent the problem, as was illustrated in the example where students who fail to obtain a minimum grade are required to register for following occurrences of a subject. It is, then, reasonable to assume that the number of registrations on any given academic year is directly influenced by the results of the previous.

The format proposed for this prediction topic was constructed on the basis of the former assumption. The predictor and response variables are presented in Table 4.1, where the entries preceded by *prev* indicate information pertaining to the subject's previous occurrence.

The variable *prev.fails* was not directly available in the dataset, and was calculated as the number of students registered in the previous occurrence minus the ones who were approved; this equates to the number of students who, theoretically, will have to be registered in the sample's edition. Similarly, the variable *registered.semester.change* was calculated as the ratio between students registered in the sample's semester and students registered in the previous occurrence. Both predictors, direct correlations between original variables, were introduced due to their potential influence in the response variable. The formula utilized to calculate the ratio is expressed by the formula:

$$registered.semester.change = \frac{registered.semester}{prev.registered.semester} \tag{4.1}$$

The predictors *code*, *curricular.year* and *semester* represent categorical variables and were, thus, converted to factors in the R language; all other predictors were represented as numbers. The categorical variables contained 37, 4 and 2 levels, respectively. Variable transformations, such as centering and scaling, were conducted on a model by model basis.

Table 4.1: Data structure used in models for non-optional curricular units

Predictors			Response		
Variables	Mean	SD	Variable	Mean	SD
code	✗	✗			
year	✗	✗			
curricular.year	✗	✗			
semester	✗	✗			
prev.registered	150.7	42.1			
prev.evaluated	127.0	29.3			
prev.approved	107.0	20.1			
prev.fails	43.7	34.9	registered	152.6	38.2
prev.evaluated.avg	13.3	2.0			
prev.evaluated.sd	2.7	0.9			
prev.approved.avg	13.9	1.3			
prev.approved.sd	2.2	0.3			
prev.registered.semester	905.7	414.5			
registered.semester	894.3	399.5			
registered.semester.change	1.0	0.1			

4.2 Results

The results, as presented by *caret*, are displayed over the following sections. All models were constructed under the same conditions in regards to environment and sampling partitions using a total of 214 samples and 15 predictor variables. The results here presented were obtained after a process of 10-fold cross-validation with each block estimated to have between 192 and 193 samples.

4.2.1 Ordinary Linear Regression

Initially, when attempting to train the predictive model with all defined predictors, R returned multiple warnings with the message *"prediction from a rank-deficient fit may be misleading"*. Rank deficiency ensues when the model lacks sufficient information in order to construct the desired model. Generally, this issue arises from two possible causes:

- More predictors than data samples. This drawback may already be present in the original data, but it may also be introduced during the initial pre-processing of categorical predictors that takes place when generating a new model. In R's implementation, categorical factors are arranged using a strategy known as *dummy coding*. The method converts each categorical predictor into multiple variables, known as *dummies*, where each variable represents a categorical level [Sta10]. For instance, the predictor *code* is transformed into the predictors *code1*, *code2*, *code3*, etc. After the initial pre-processing stage, if the total number of predictors surpasses the number of data samples, a deficiency will occur.

- Collinear variables do not increase the model's numerical rank. Essentially, this issue refers to predictors that do not introduce new information.

When reviewing the original predictors in further detail, the three categorical variables (*code*, *curricular.year* and *semester*) reveal a combined total of 43 levels. This means that, even when considering dummy coding, the total number of samples surpasses the number of predictors. Additional experiments later revealed that the warnings occurred whenever the predictor *code* was used in conjunction with *curricular.year* or *semester*. It was also recognized that the addition of the later two predictors to a model with *code* did not improve the cross-validated RMSE. Consequently, the predictors *curricular.year* and *semester* were removed from following phases.

Table 4.2: Results for all possible combinations of predictors in linear regression models for non-optional curricular units

N Predictors	Combinations	Min RMSE	Max RMSE	Mean RMSE
4	715	9.922	32.789	18.107
5	1287	9.754	31.211	15.982
6	1716	9.764	30.376	14.278
7	1716	9.765	28.125	12.934
8	1287	9.783	25.845	11.909
9	715	9.825	22.773	11.164
10	286	9.870	20.756	10.652
11	78	9.921	12.342	10.323
12	13	9.973	11.243	10.130
13	1	10.035	10.035	10.035

Additional experimentation found that, on average, the construction of a complete linear regression model never exceeded 0.6 seconds. As such, it was both possible and viable to train and evaluate models for all possible combinations of predictors. The results, in the form of cross-validated RMSE, are summarized in Table 4.2. Each entry refers to all possible combinations of predictors using n predictors. The min and max RMSE columns pertain to the worst and best models, respectively, generated with n variables. As evidenced, although the RMSE average is higher in models with fewer predictors, the minimum corresponding to the models with lower RMSE is maintained and even improved.

Table 4.3 presents the predictors utilized in the best models per number of predictors. It is clear that the first four predictors (*code*, *prev.fails*, *registered.semester*, *year*) are always maintained and, generally, an extra predictor is introduced and then maintained. After five predictors, the best models exhibit a higher RMSE value in the orders of 10^{-2}.

Table 4.3: Predictors used in the best linear regression models per number of predictors for non-optional curricular units

	Predictors Used								
Predictor	4	5	6	7	8	9	10	11	12
code	✓	✓	✓	✓	✓	✓	✓	✓	✓
prev.fails	✓	✓	✓	✓	✓	✓	✓	✓	✓
registered.semester	✓	✓	✓	✓	✓	✓	✓	✓	✓
year	✓	✓	✓	✓	✓	✓	✓	✓	✓
registered.semester.change		✓	✓	✓	✓			✓	✓
prev.evaluated			✓		✓	✓	✓	✓	✓
prev.approved				✓	✓	✓	✓	✓	✓
prev.registered				✓	✓	✓	✓	✓	✓
prev.approved.sd						✓	✓	✓	✓
prev.registered.semester						✓	✓	✓	✓
prev.evaluated.sd							✓	✓	✓
prev.evaluated.avg									✓
RMSE	9.922	9.754	9.764	9.765	9.783	9.825	9.870	9.921	9.973

It should be noted, however, that differences in these orders are highly influenced by the cross-validation process, and it is quite probable that differently randomized partitions would result in different, albeit similar, results. The final model, selected on the basis of RMSE and simplicity, used five predictors. Its results are presented in Table 4.4.

Table 4.4: Results for the linear regression model selected for non-optional curricular units

RMSE	R^2
9.754	0.931

4.2.2 Partial Least Squares

The process conducted for the PLS models was similar to the one applied to ordinary linear regression, where an exhaustive search of all combinations of predictors was performed. Prior to the training phase, all predictors were centered and scaled.

For each generated model, the number of components to keep (parameter *ncomp*) was varied between 1 and $npredictors - 1$. This implies that, for a given model with nine predictors, eight different models are generated. The results are summarized in Table 4.5. It should be noted that the number of combinations displayed does not consider that, for each model, additional models are constructed for the *ncomp* variation. As with ordinary linear regression, the RMSE average is higher on models with fewer predictions. However, unlike with the previous, the best results are found in intermediate levels.

Table 4.5: Results for all possible combinations of predictors in PLS models for non-optional curricular units

N Predictors	Combinations	Min RMSE	Max RMSE	Mean RMSE
4	715	10.353	32.731	19.114
5	1287	10.104	30.929	16.709
6	1716	10.072	30.043	14.861
7	1716	9.918	28.111	13.398
8	1287	9.814	25.743	12.273
9	715	9.856	22.685	11.439
10	286	9.891	20.689	10.848
11	78	9.883	14.145	10.454
12	13	9.960	11.157	10.199
13	1	10.017	10.017	10.017

Table 4.6 presents the predictors utilized in the best models per number of predictors. Similarly, four predictors are always maintained, although *registered.semester.change* replaces *code*. The five most common predictors are the same in both OLS and PLS. Overall, there is more variation regarding the variables maintained as the number of predictors increases in PLS.

Table 4.6: Predictors used in the best PLS models per number of predictors for non-optional curricular units

	Predictors Used								
Predictor	4	5	6	7	8	9	10	11	12
code		✓	✓	✓	✓	✓	✓	✓	✓
prev.fails	✓	✓	✓	✓	✓	✓	✓	✓	✓
registered.semester	✓	✓	✓	✓	✓	✓	✓	✓	✓
year	✓	✓	✓	✓	✓	✓	✓	✓	✓
registered.semester.change	✓	✓	✓	✓	✓	✓	✓	✓	✓
prev.evaluated					✓		✓	✓	
prev.approved			✓		✓	✓	✓	✓	✓
prev.registered								✓	✓
prev.approved.sd								✓	✓
prev.registered.semester						✓	✓		✓
prev.evaluated.sd								✓	✓
prev.evaluated.avg				✓	✓	✓	✓		✓
prev.approved.avg				✓		✓	✓	✓	✓
RMSE	10.353	10.104	10.072	9.918	9.814	9.856	9.891	9.883	9.960

The final model, selected on the basis of RMSE, used eight predictors and five components. Its results are presented in Table 4.7.

Table 4.7: Results for the PLS model selected for non-optional curricular units

RMSE	R^2
9.953	0.931

4.2.3 Elastic Net

The process applied to the generation of elastic net models had the predictors centered and scaled prior to the training phase. It should be noted that *caret's* implementation of this model is not capable of handling categorical values and, therefore, the predictors *code*, *curricular.year* and *semester* were manually transformed using the dummies strategy.

The weight decay (parameter *lambda*) was varied from a list with the values 0, 0.001, 0.01 and 0.1, while the fraction of the full solution (parameter *fraction*) was taken from a sequence of twenty equally separated values between 0.05 and 1. Resampling results for the generated models are presented in Table 4.8. Figure 4.1 illustrates how the RMSE behaves in function of the parameters.

Table 4.8: Sample of the results for the elastic net models for non-optional curricular units

lambda	fraction	RMSE	R^2
0.000	0.05	13.828	0.916
0.000	0.10	11.164	0.919
0.000	0.15	10.133	0.925
0.000	0.20	10.014	0.927
...	...	...	...
0.010	0.40	9.972	0.933
...	...	...	...
0.100	1.00	11.362	0.920

Generally, the error converged to a similar minimum regardless of the weight decay, with lower weights converging faster. RMSE was used to select the optimal model, with final values for *fraction* and *lambda* set at 0.4 and 0.01, respectively.

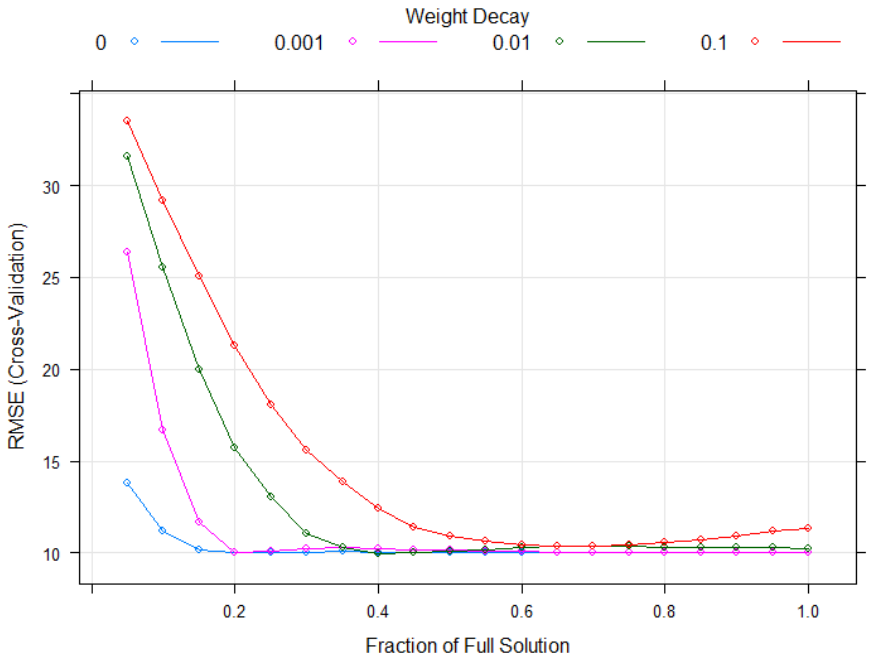

Figure 4.1: Elastic net model variations for non-optional curricular units

4.2.4 Multivariate Adaptive Regression Splines

Regression models based on MARS were generated with no predictor transformations. The number of terms maintained (parameter *nprune*) was varied between 2 and 25, while the product degree

(parameter *degree*) was held constant at a value of 1. Resampling results for the generated models are presented in Table 4.8. Figure 4.2 illustrates how the RMSE behaves in function of the parameters.

Table 4.9: Sample of the results for the MARS models for non-optional curricular units

nprune	RMSE	R^2
2	16.913	0.806
3	13.357	0.874
4	11.714	0.902
5	11.231	0.914
...	...	...
16	9.006	0.945
...	...	...
25	9.197	0.943

The error was found to be reduced as more terms were maintained, stabilizing after 18 terms. RMSE was used to select the optimal model, with *nprune* set at 16.

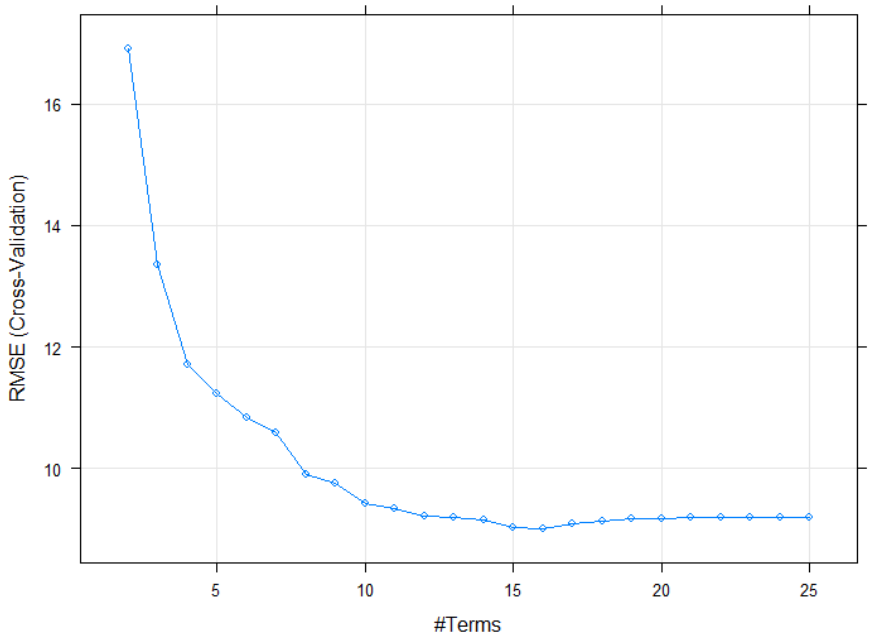

Figure 4.2: MARS model variations for non-optional curricular units

4.2.5 Support Vector Machines

The SVM models constructed for this topic used a radial basis function as the kernel. Prior to the training phase, all predictors were centered and scaled. As with the models generated using the elastic net strategy, categorical variables were manually transformed into dummies.

Experimentation found that the removal of some variables improved the model fit. As a result, predictors were individually subtracted from the model based on the variable importance identified by *caret* until the resampled RMSE increased. Table 4.10 presents the predictors removed in each iteration. At the conclusion of the feature selection process, the following predictors had been removed: *code*, *prev.approved*, *prev.approved.sd*, *prev.evaluated* and *prev.evaluated.sd*.

Table 4.10: Predictors removed in the SVM models for non-optional curricular units

	Predictors Removed						
Predictor	0	1	2	3	4	5	6
code		✗	✗	✗	✗	✗	✗
prev.approved			✗	✗	✗	✗	✗
prev.approved.sd				✗	✗	✗	✗
prev.evaluated					✗	✗	✗
prev.evaluated.sd						✗	✗
year							✗
RMSE	12.322	11.423	10.593	10.351	9.927	9.860	11.205

For each configuration of variables, the cost (parameter C) was taken from a sequence of fifteen values between 0.25 and 4096 with each value being equal to the double of the previous, while the scaling factor (parameter *sigma*) was held constant at the value of 0.07540113, as set by *caret*. Resampling results for the models generated with the final combination of predictor variables selected are presented in Table 4.11. Figure 4.3 illustrates how the RMSE behaves in function of the parameters.

Table 4.11: Sample of the results for the SVM models for non-optional curricular units

C	RMSE	R^2
0.25	16.123	0.852
0.50	13.330	0.895
1.00	11.415	0.918
2.00	10.393	0.929
4.00	9.860	0.933
8.00	9.887	0.931
...	...	...
4096.00	11.085	0.913

The error stabilized once the cost surpassed 128, and remained constant throughout. RMSE was used to select the optimal model, with the cost set at 4.

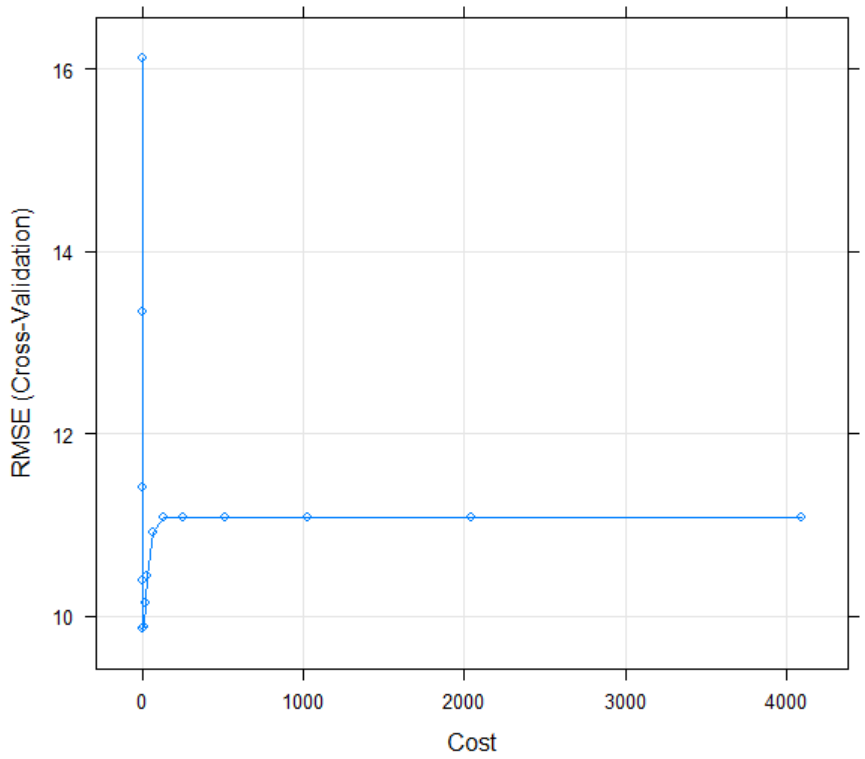

Figure 4.3: SVM model variations for non-optional curricular units

4.2.6 Neural Networks

The process applied to the generation of the neural networks was based on an ensemble strategy known as model averaging, where multiple models using the same parameters are fit using different random number seeds [MWK+16]. Afterwards, predictions are averaged from the resulting models. Each ensemble was constructed from 5 neural networks (note that this is *caret*'s default, and cannot be configured). Prior to the training phase, all predictors were centered and scaled.

The weight decay (parameter *decay*) was tested from a list with the values 0.001, 0.01 and 0.1, while the number of hidden units (parameter *size*) was varied from a sequence initiated at 1 and increased two units until 27. Resampling results for the generated models are presented in Table 4.12. Figure 4.4 illustrates how the RMSE behaves in function of the parameters.

The parameters presented resulted in a total computation time of over 60 minutes. For comparison, all the other models generated for this topic required a combined time of less than 3 minutes.

Table 4.12: Sample of the results for the neural network models for non-optional curricular units

decay	size	RMSE	R^2
0.001	1	11.653	0.912
0.001	3	12.439	0.895
0.001	5	15.387	0.863
0.001	7	15.776	0.836
...	...	...	...
0.010	1	9.961	0.931
...	...	...	...
0.100	27	NaN	NaN

The final model, selected on the basis of RMSE, used 1 hidden unit and had the weight decay set at 0.01.

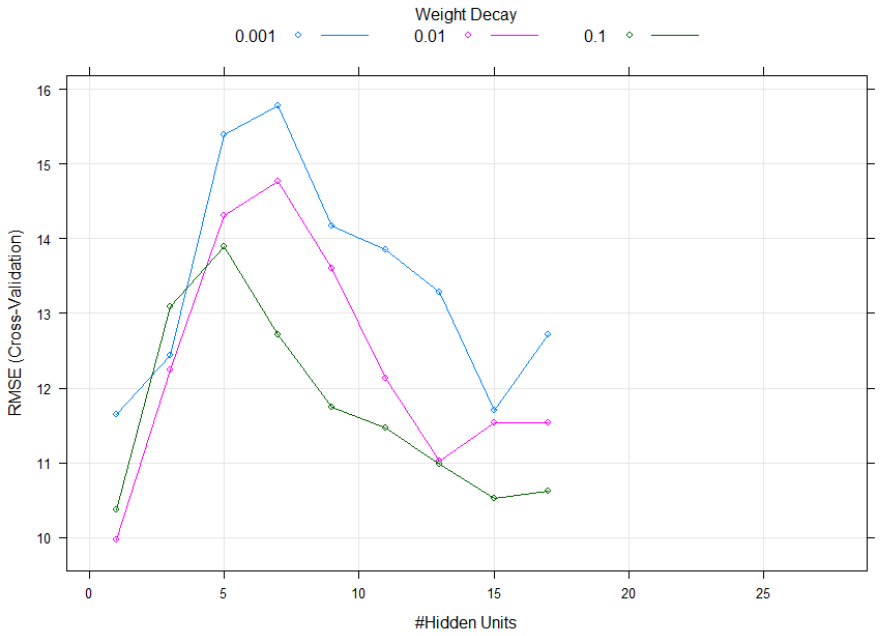

Figure 4.4: Neural network model variations for non-optional curricular units

4.2.7 k-Nearest Neighbors

Regression models based on k-NN had the predictors centered and scaled prior to the training phase. Categorical variables were manually transformed using the dummies strategy. Initial experiments with all predictors revealed an average error of over 18, which indicated the presence of one or more dependent variables that did not contribute significantly to the outcome. Additional tests, conducted on the basis of variable importance and the combinations of variables obtained

with OLS and PLS, discovered improved results with the removal of the following predictors: *code*, *curricular.year*, *prev.evaluated*, *prev.evaluated.avg*, *prev.evaluated.sd* and *prev.approved.sd*.

The number of neighbors averaged (parameter k) was varied between 1 and 15, using the Euclidean distance. Resampling results for the generated models are presented in Table 4.13. Figure 4.5 illustrates how the RMSE behaves in function of the parameters.

Table 4.13: Sample of the results for the k-NN models for non-optional curricular units

k	RMSE	R^2
1	14.823	0.855
2	12.629	0.895
3	13.271	0.894
4	13.319	0.894
5	13.647	0.895
...	...	...
15	17.442	0.836

In all instances with more than two neighbors observed, the average error increased as more neighbors were added. RMSE was used to select the optimal model, with the number of neighbors set at 2.

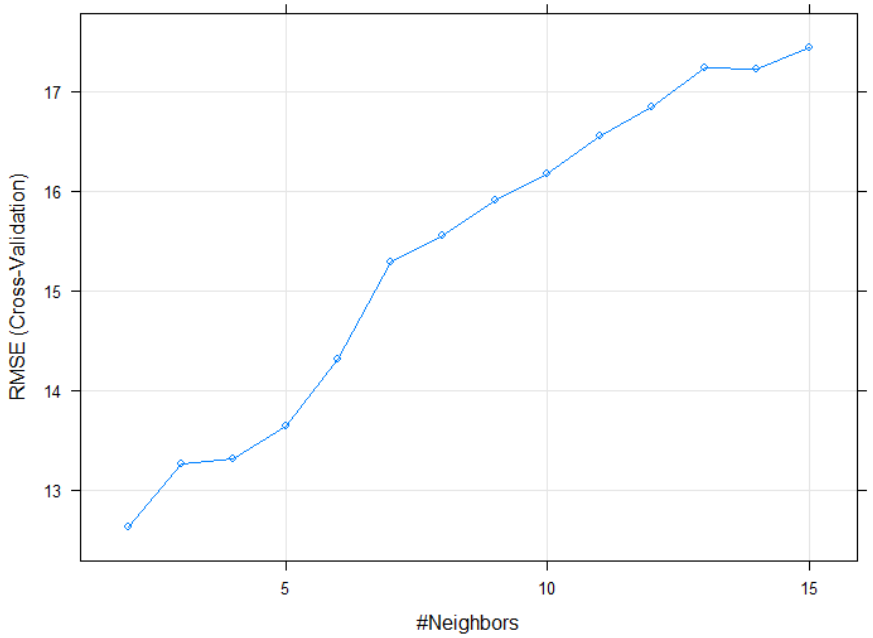

Figure 4.5: k-NN model variations for non-optional curricular units

4.2.8 Basic Regression Trees

The process conducted for the generation of the models here presented, based on the Classification and Regression Tree (CART) methodology, did not require any variable transformations. This is also the case in the majority of tree-based models, as will be evidenced over the following sections.

The complexity parameter utilized in the pruning phase (parameter *cp*) was varied from a sequence of thirty equally separated values between 0.00 and 0.57. Resampling results for the generated models are presented in Table 4.14. Figure 4.6 illustrates how the RMSE behaves in function of the parameters.

Table 4.14: Sample of the results for the CART models for non-optional curricular units

cp	RMSE	R^2
0.000	16.432	0.843
0.020	17.777	0.798
0.039	19.845	0.740
0.059	20.029	0.737
0.079	20.029	0.737
..	...	...
0.570	34.357	0.512

As evidenced, the increase of the complexity parameter resulted in a higher average error. RMSE was used to select the optimal model, with the complexity set at 0.0. The resulting tree exhibited a total of 29 nodes, with 15 leaves.

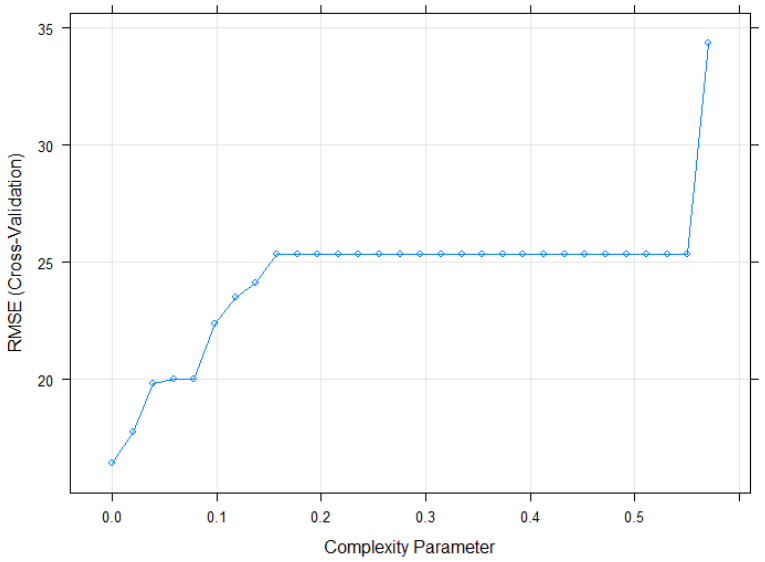

Figure 4.6: CART model variations for non-optional curricular units

4.2.9 Conditional Inference Trees

As with the trees illustrated in the previous section, models constructed for this topic did not require predictor transformations. For evaluation purposes, the $1-p$-value statistical threshold (parameter *mincriterion*) was varied between 0.01 and 0.99. Resampling results for the generated models are presented in Table 4.15. Figure 4.7 illustrates how the RMSE behaves in function of the parameters.

Table 4.15: Sample of the results for the conditional inference tree models for non-optional curricular units

mincriterion	RMSE	R^2
0.010	14.939	0.860
0.119	14.942	0.859
0.228	15.104	0.857
0.337	15.076	0.859
0.446	15.077	0.858
...	...	...
0.990	17.007	0.822

As observed in most instances, the increase of the *mincriterion* lead to a higher average error. The final model, selected using the RMSE as the evaluation metric, had the *mincriterion* set at 0.1.

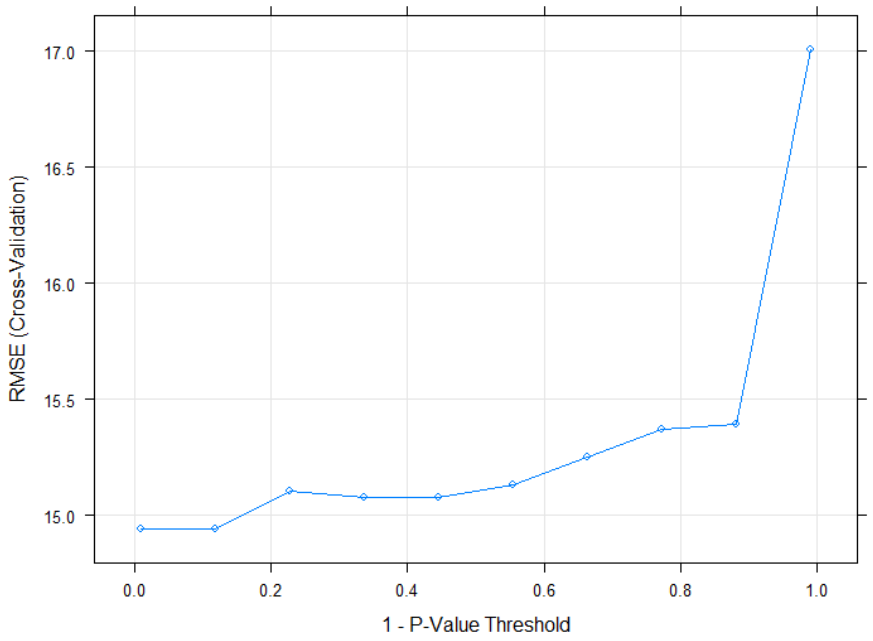

Figure 4.7: Conditional inference tree model variations for non-optional curricular units

4.2.10 Model and Rules Trees

The modeling approach selected for this topic is based on the M5 algorithm, with no predictor transformations applied. All combinations of pruning and smoothing processes were tested. Resampling results for the generated models are presented in Table 4.16. Figure 4.8 illustrates how the RMSE behaves in function of the parameters.

Table 4.16: Results for the M5 models for non-optional curricular units

pruned	smoothed	rules	RMSE	R^2
Yes	Yes	Yes	10.411	0.926
Yes	Yes	No	10.491	0.925
Yes	No	Yes	10.466	0.925
Yes	No	No	10.599	0.923
No	Yes	Yes	11.220	0.912
No	Yes	No	10.348	0.927
No	No	Yes	14.836	0.848
No	No	No	14.978	0.851

The final model, selected on the basis of RMSE, used a smoothing process and no pruning.

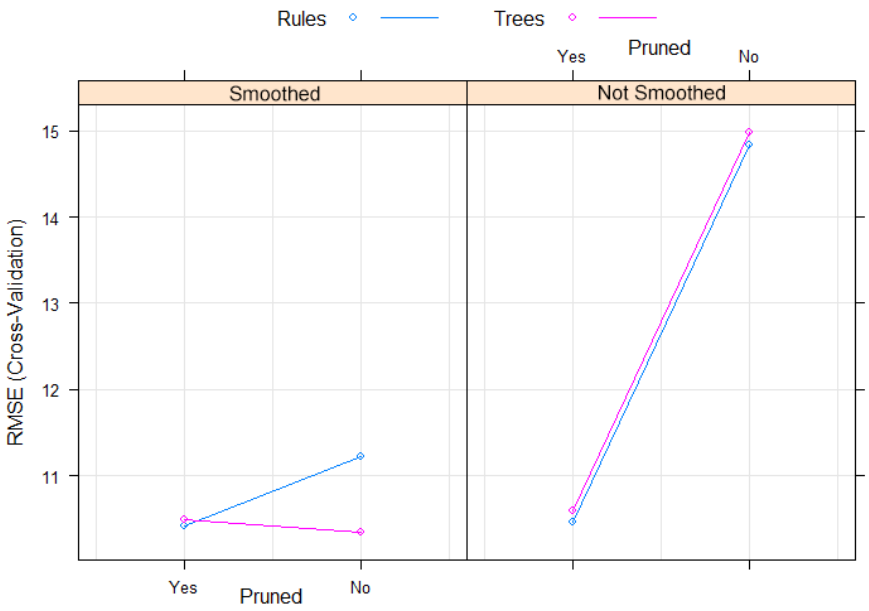

Figure 4.8: M5 model variations for non-optional curricular units

4.2.11 Cubist

The process conducted for the generation of the Cubist models here presented did not require any variable transformations. The number of committees and neighbors, represented by parameters of identical names, were taken from lists with the values 1, 5, 10, 50, 75 and 0, 1, 3, 5, 7, 9, respectively. Resampling results for the generated models are presented in Table 4.17. Figure 4.9 illustrates how the RMSE behaves in function of the parameters.

Table 4.17: Sample of the results for the cubist models for non-optional curricular units

committees	neighbors	RMSE	R^2
1	0	10.780	0.916
1	1	12.061	0.896
1	3	10.898	0.921
1	5	10.359	0.926
...	...	...	...
50	5	9.353	0.942
...	...	...	...
75	9	9.453	0.939

The final model, selected using the RMSE as the evaluation metric, utilized 50 committees and 5 neighbors.

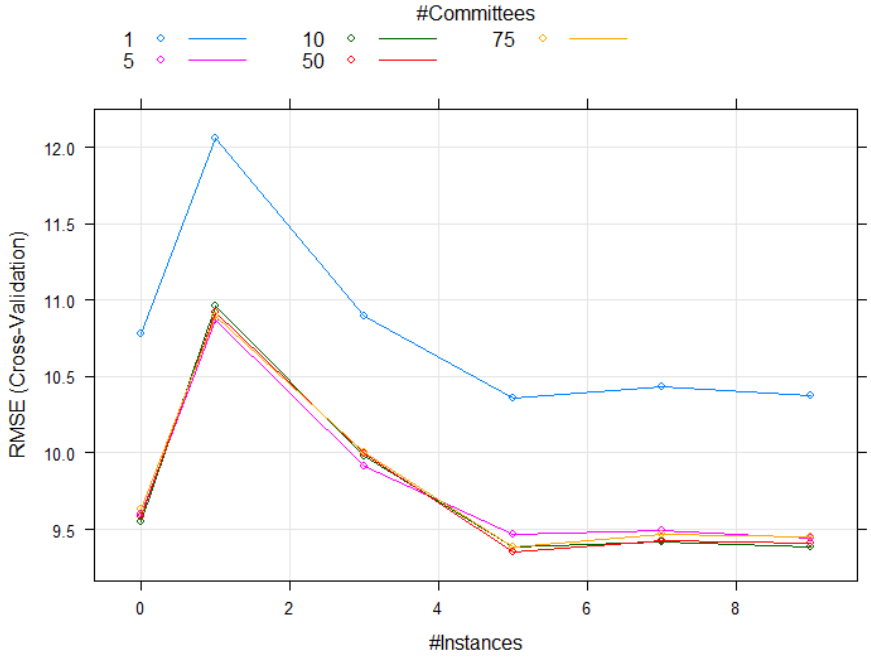

Figure 4.9: Cubist model variations for non-optional curricular units

4.2.12 Bagged Trees

The base tree models utilized in the bagged ensemble constructed for this topic were based on the CART methodology. The final ensemble was constructed from 25 trees (note that this is *caret*'s default, and cannot be configured). Resampling results for the generated model are presented in Table 4.18.

Table 4.18: Results for the bagged tree models for non-optional curricular units

RMSE	R^2
14.102	0.877

4.2.13 Random Forests

Regression models based on random forests were generated with no predictor transformations. The number of randomly selected predictors at each split (parameter *mtry*) was varied between 2 and the number of predictors. Resampling results for the generated models are presented in Table 4.19. Figure 4.10 illustrates how the RMSE behaves in function of the parameters.

Table 4.19: Sample of the results for the random forest models for non-optional curricular units

cp	RMSE	R^2
2	13.346	0.900
3	12.703	0.906
4	12.436	0.908
5	12.231	0.910
...	...	...
10	11.880	0.909
...	...	...
15	12.119	0.904

Note that the results here presented refer to a forest with 1000 trees. Additional tests with 500 trees and increments of 500 up to a total of 5000 trees yielded no improvements. RMSE was used to select the optimal model, with *mtry* set at 10 predictors.

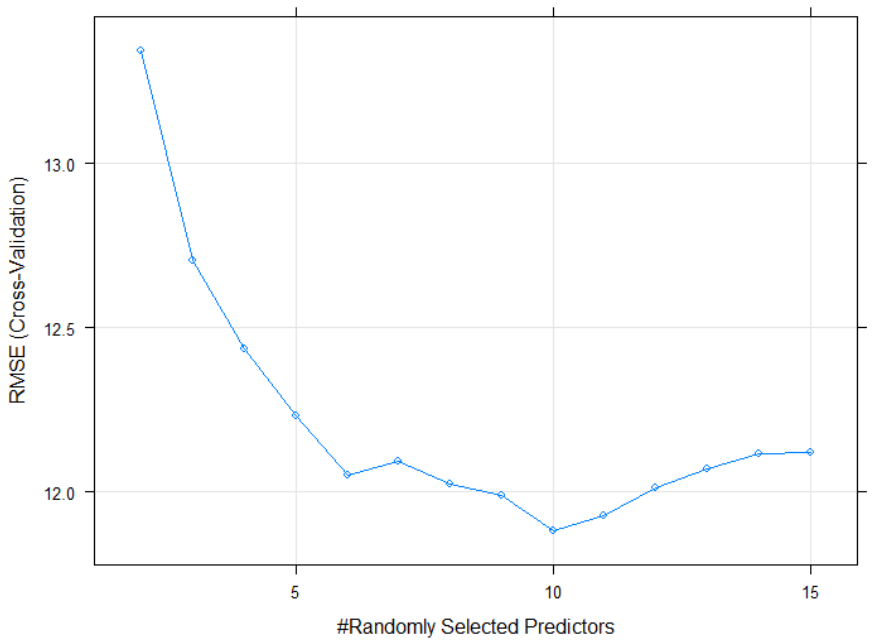

Figure 4.10: Random forest model variations for non-optional curricular units

4.2.14 Boosting

The process applied to the generation of the boosting models was based on the *gbm* package, which closely follows Friedman's Gradient Boosting Machine [Sou15]. The process analyzed did not require variable transformations. The number of iterations, or trees (parameter *n.trees*), was varied from a sequence initiated at 100 and increased by 100 units until 5000, while the maximum tree depth (parameter *interaction.depth*) was taken from a list with the values 1, 3, 5, 7, 9 and 11. The regularization penalty (parameter *shrinkage*) was tested with 0.01 and 0.1, and the minimum size per terminal node (parameter *n.minobsinnode*, corresponding to the minimum number of observations per terminal node) was held constant at a value of 1. Resampling results for the generated models are presented in Table 4.20. Figure 4.11 illustrates how the RMSE behaves in function of the parameters.

Table 4.20: Sample of the results for the boosting models for non-optional curricular units

shrinkage	interaction.depth	n.trees	RMSE	R^2
0.01	1	100	25.257	0.792
0.01	1	200	19.426	0.843
0.01	1	300	16.105	0.866
0.01	1	400	14.321	0.879
...	...	...	...	...
0.01	3	4900	10.220	0.928
...	...	...	...	...
0.10	11	5000	11.323	0.915

Additional tests with an increased number of observations per terminal node showed no signs of improvement. RMSE was used to select the optimal model, with final values for *n.trees*, *interaction.depth* and *shrinkage* set at 4900, 3 and 1, respectively.

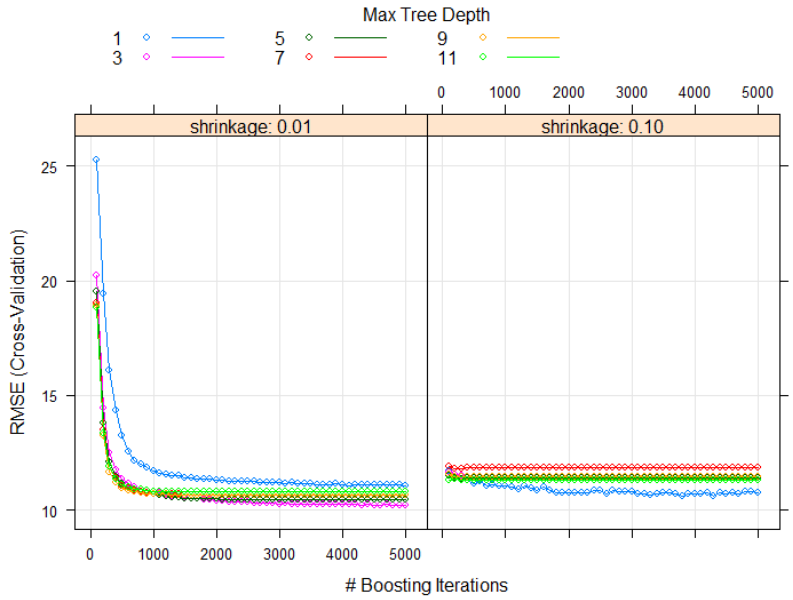

Figure 4.11: Boosting model variations for non-optional curricular units

4.2.15 Aggregated Results

Overall, all the models presented in this section proved relatively successful at predicting the number of students registered per non-optional curricular unit. Table 4.21 presents a brief summary, ordered by RMSE, of the resampling results obtained with 10-fold cross-validation for the best models constructed for each regression algorithm. The results obtained for the first half of

the models are fairly competitive, with MARS pulling slightly ahead. Linear regression models placed similarly, indicating that the prediction topic can be accurately described by a linear combination of the predictor variables. Models reliant on predictions constructed from sample averages, as were the cases with k-NN and several tree-based models, demonstrated less adaptability and resulted in a worse fit.

Table 4.21: Summary of the results for all models tested for non-optional curricular units

	RMSE	R^2
MARS	9.006	0.9448
Cubist	9.353	0.9421
Linear Regression	9.754	0.9313
SVM	9.860	0.9332
Partial Least Squares	9.953	0.9319
Neural Networks	9.961	0.9307
Elastic Net	9.972	0.9326
Boosting	10.220	0.9284
Model Trees	10.350	0.9269
Random Forests	11.880	0.9095
k-NN	12.630	0.8954
Bagged CART	14.100	0.8768
Cond Inference Trees	14.940	0.8595
CART	16.430	0.8427

An additional test, portraying a more accurate scenario, was also conducted. Generally, as the predictions per unit need to be calculated before the semester has begun, data for other subjects in the same semester is not yet available. The resampling process utilized to estimate and compare model performance does not consider the former factor and is, thus, incapable of capturing all the nuances of a real-world scenario. The test scenario partitioned the data into two blocks: a training set, composed of the academic years of 2009/2010 to 2014/2015, and a test set pertaining to the academic year of 2015/2016. The train and test sets were built from 177 and 37 samples, respectively.

The three best regression models previously obtained were compared between themselves and against other estimates. These estimates, which also considered the previously established data partitions, are briefly described below:

- **Naive Average**: assumes each prediction to be equal to the mean of the observations in the training set.

- **Naive Previous**: calculates the prediction for a given curricular unit as the mean of all its previous occurrences.

- **Informed Previous**: calculates the prediction for a given curricular unit as the number of students registered in the previous occurrence added to half of the difference between the previous two editions. This is currently one of the main strategies employed by MIEIC's management body, and aims to capture the most recent trend in the data. The formula for this calculation can be expressed as follows:

$$\hat{y}_t = y_{t-1} + \frac{y_{t-1} - y_{t-2}}{2} \tag{4.2}$$

  where $\hat{y}$ represents the prediction value, y represents an observation, and t refers to the year of a given observation.

The results for the test case are presented in Table 4.22. As depicted, all predictive models managed to surpass the other estimates, even reducing the average error by more than half under some circumstances. When considering the mean response of 152.6 students registered per curricular unit, the models achieve an average error of around 5%. As a unit of reference, units are typically composed of classes with 20 to 25 students.

It is also fundamental to mention that, while the regression models are expected to maintain the average evidenced by the resampling results, the same does not apply to the estimates described. The informed estimate, for instance, displays a RMSE of 25.427 and 34.005 when applied to the academic years of 2014/2015 and 2013/2014, respectively.

Table 4.22: Comparison between the three best regression models constructed for non-optional curricular units and other estimates

Prediction	RMSE
MARS	8.854
Cubist	9.464
Linear Regression	6.311
Naive Average	28.496
Naive Previous	20.004
Informed Previous	15.114

4.3 Experiments

After concluding the modeling phase, experiments were mainly focused on the construction of heterogeneous ensembles. These ensembles differ from the ones already evaluated, such as random forests or bagged trees, in the fact that the individual models that compose the ensemble are based on various types of regression algorithms. The process here presented analyzes ensembles built from combinations of the models that obtained the best results in the previous step. All the results examined in this section were derived after a process of 10-fold cross-validation.

4.3.1 Ensemble: Generalized Linear Model

The initial ensemble was constructed via a Generalized Linear Model (GLM) that describes a linear combination of three of the models previously described. The approach is based on the R package *caretEnsemble* which, as of the moment of this publication, does not support individual models with distinct predictor variables or different variable transformations. To account for this, models such as the linear regression presented in Section 4.2 were removed from the experiment. The three models utilized, based on their resampled RMSE, were MARS, cubist and boosting.

Results for the ensemble are presented in Table 4.23. Although the differences are minimal, the new combined model surpasses any individual algorithm, achieving the best RMSE and R^2 values reviewed thus far. When evaluated under the test scenario's conditions, the ensemble model achieves a RMSE of 7.814.

Table 4.23: Results for the GLM ensemble constructed for non-optional curricular units

RMSE	R^2
8.868	0.946

4.3.2 Ensemble: Stacking

The approach selected for this topic is based on a stacking method which utilizes an extra algorithm to combine the predictions of various other models, generating an ensemble. The ensembling algorithm may be seen as a *meta* model. The individual regression models selected for this topic were MARS, cubist and boosting, which were then combined via random forests, cubist and boosting. Note that, once more, ordinary linear regression was removed from the experiment due to *caretEnsemble*'s limitations.

Table 4.24: Summary of the results for the stacking ensembles constructed for non-optional curricular units

	RMSE	R^2
Cubist	9.271	0.941
Random Forests	9.948	0.936
Boosting	12.791	0.906

Table 4.24 presents a summary of the results obtained with the stacking models. While the cubist ensemble demonstrates a slight improvement from the individual models which compose the ensemble, it remains inferior to the ensemble illustrated in the previous experiment.

4.3.3 Ensemble: Bagging

The final ensemble was approached with a bagging process that combined the models of MARS, cubist and ordinary linear regression. The ensembling algorithm did not rely on any package, which allowed for the inclusion of the previously analyzed ordinary linear regression model. A simplified version of the bagging algorithm utilized, adapted from Kuhn and Johnson [KJ13], is presented in Algorithm 1.

Algorithm 1 Bagging Algorithm

1: **for** $i \leftarrow 1, iterations$ **do**
2: Generate a bootstrap sample of the training data
3: Train a model of each type on this sample
4: **end for**

Each individual model is responsible for generating predictions, which are then averaged to produce the bagged ensemble's prediction. The final ensemble utilized 200 iterations, for a total of 600 individual predictive models, resulting in a RMSE of 10.037. Altogether, the bagged model falls short of the ensembles depicted in previous experiences.

4.4 Conclusions

Based on the resampling procedure selected for model evaluation and performance estimation, it can be inferred that the best regression method constructed for this topic is one based on a **GLM ensemble**. The ensemble, presented in detail in Section 4.3, is composed of a combination of the MARS, cubist and boosting models defined in Section 4.2. After a resampling process of 10-fold cross-validation, the ensemble is estimated to have a RMSE of 8.868, surpassing any existing alternative currently used by the course's administration.

When reviewing the agglomeration of models constructed for the prediction topic, it can be inferred that most models would prove capable of successfully predicting the number of registrations for future occurrences. Overall, the results presented are primarily positive, expressing the advantages of complete predictive models over simple estimates and extrapolations.

Chapter 5

Number of Students Enrolling in Optional Curricular Units

This chapter presents the results obtained with the predictive models constructed for the prediction of the number of students registered in optional curricular units per semester. Section 5.1 describes, in detail, the data structure selected for the topic, while Section 5.2 illustrates the process applied to and results obtained by each individual model. Additional experiments are discussed in Section 5.3. Lastly, Section 5.4 concludes on which model provides a better fit to the problem.

5.1 Experimental Setup

The original dataset constructed for this prediction topic included data relative to all curricular years, per semester, from the academic years of 2009/2010 to 2015/2016. Initial analysis revealed that, in the majority of the samples, there were no students registered in optional curricular units. In MIEIC, students are only allowed to select optional units after that their fourth curricular year. In fact, the selection process only occurs during the first and second semesters of the fourth curricular year, and the first semester of the fifth, as the second semester is typically reserved for the dissertation. As such, it is not necessary to predict a response for other semesters, and the corresponding entries were removed from the dataset.

The filtering process, despite resulting in a consistent dataset with no irrelevant observations, also shortened the original data from 71 to 21 entries. This reduced number of observations may impact the predictive models' performance, and is considered a direct influence in the results obtained.

The independent and dependent variables are presented in Table 5.1. The predictors *curricular.year* and *semester* were converted to factors with 2 levels; all other predictors were represented as numbers. Variable transformations were conducted on a model by model basis.

Table 5.1: Data structure used in models for optional curricular units per semester

Predictors			Response		
Variables	Mean	SD	Variable	Mean	SD
year	✗	✗	optionals.registered	205.8	126.4
curricular.year	✗	✗			
semester	✗	✗			
students	570.1	81.2			
mobility.in	14.7	8.3			
mobility.out	16.4	10.3			

It is worth noting that the response variable's mean may be somewhat misleading. In the first semester of the fourth curricular year, the average number of students enrolling in optional units is much fewer, with an average of 38.4.

5.2 Results

The results, as presented by *caret*, are displayed over the following sections. All models were constructed under the same conditions in regards to environment and sampling partitions using a total of 21 samples and 6 predictor variables. The results here presented were obtained after a process of 10-fold cross-validation with each block estimated to have between 18 and 19 samples.

The selection of the models depicted in this topic was mainly grounded on performance capabilities. Some regression models, such as PLS, were included due to their positive results in Chapter 4.

5.2.1 Ordinary Linear Regression

Regression models based on ordinary linear regression were generated with no predictor transformations. As there were no tuning parameters, the results are presented directly in Table 5.2.

Table 5.2: Results for the linear regression model for optional curricular units per semester

RMSE	R^2
41.865	0.980

5.2.2 Partial Least Squares

The process applied to the generation of PLS models had the predictors centered and scaled prior to the training phase. For each generated model, the number of components to keep (parameter *ncomp*) was varied between 1 and $npredictors - 1$. Resampling results for the generated models are presented in Table 5.3. Figure 5.1 illustrates how the RMSE behaves in function to *ncomp*.

Table 5.3: Results for the PLS models for optional curricular units per semester

ncomp	RMSE	R^2
1	113.362	0.966
2	107.541	0.959
3	77.818	0.978
4	58.885	0.941
5	44.234	0.943

As evidenced, the average error is greatly reduced with the addition of more components. The final model, selected on the basis of RMSE, utilized 5 components.

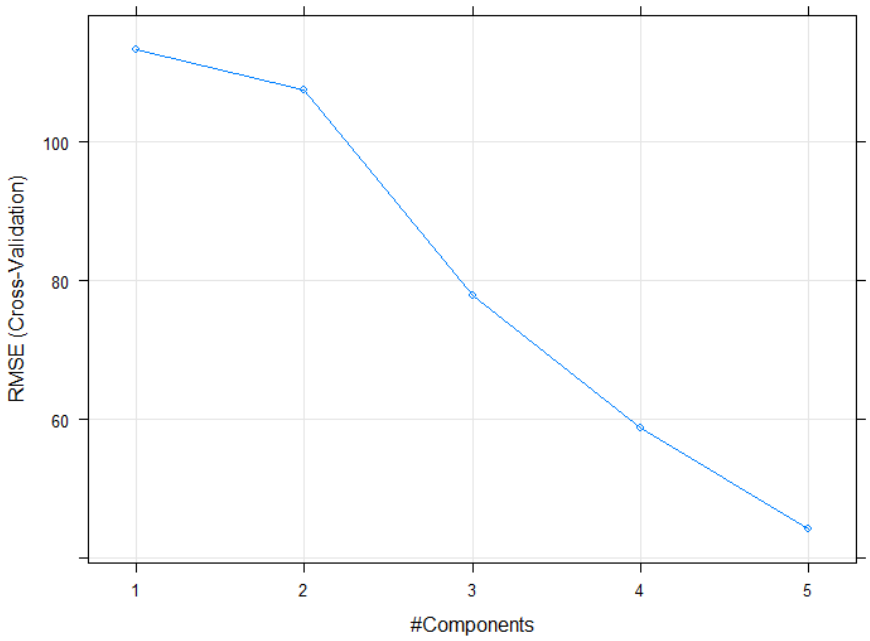

Figure 5.1: PLS model variations for optional curricular units per semester

5.2.3 Multivariate Adaptive Regression Splines

The process conducted for the generation of the MARS models here presented did not require any variable transformations. The number of terms maintained (parameter *nprune*) was varied between 2 and 25, while the product degree was held constant at a value of 1. Resampling results for the generated models are presented in Table 5.4. Figure 5.2 illustrates how the RMSE behaves in function to *ncomp*.

Table 5.4: Sample of the results for the MARS models for optional curricular units per semester

nprune	RMSE	R^2
2	143.102	1.0000
3	35.526	1.0000
4	33.254	0.9998
5	33.254	0.9998
6	33.254	0.9998
...	...	...
25	33.254	0.9998

The error was found to converge after 4 terms. RMSE was used to select the optimal model, with *nprune* set at 4.

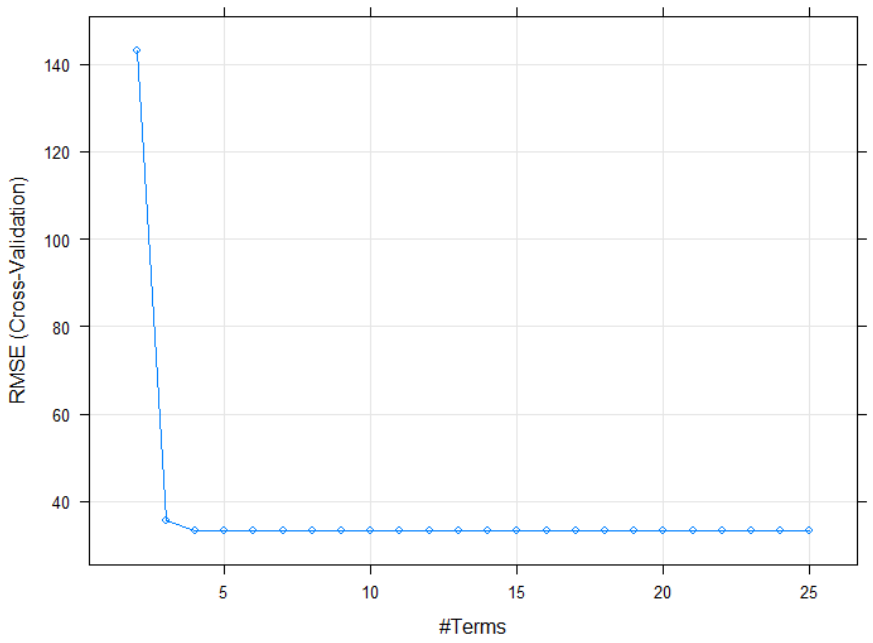

Figure 5.2: MARS model variations for optional curricular units per semester

5.2.4 Support Vector Machines

The SVM models constructed for this topic used a radial basis function as the kernel. Prior to the training phase, all predictors were centered and scaled, and categorical variables were manually transformed via dummy coding. This strategy is described, in detail, in Section 4.2. Initial experimentation found that the removal of some variables improved the model fit. As a result, predictors were individually subtracted from the model based on the variable importance identified

by *caret* until the resampled RMSE increased. Both variables pertaining to *mobility*, *mobility.in* and *mobility.out*, were removed.

The cost (parameter C) was taken from a sequence of fifteen values between 0.25 and 4096 with each value being equal to the double of the previous, while the scaling factor (parameter sigma) was held constant at the value of 0.5812391, as set by caret. Resampling results for the generated models are presented in Table 5.5. Figure 5.3 illustrates how the RMSE behaves in function of the parameters.

Table 5.5: Sample of the results for the SVM models for optional curricular units per semester

C	RMSE	R^2
0.25	91.003	0.9994
0.50	79.603	0.9999
1.00	56.398	0.9997
2.00	51.922	0.9998
4.00	51.922	0.9998
...	...	...
4096.00	51.922	0.9998

The error stabilized once the cost surpassed 2, and remained constant throughout. RMSE was used to select the optimal model, with the cost set at 2. For comparison, it is worth noting that the inclusion of both *mobility* variables displayed a RMSE of 70.676.

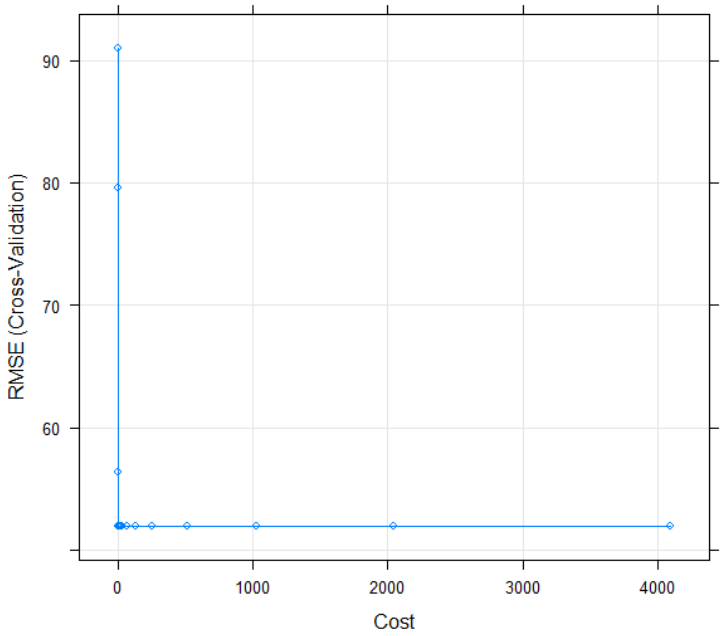

Figure 5.3: SVM model variations for optional curricular units per semester

5.2.5 Neural Networks

The process applied to the generation of the neural networks was based on the ensembling strategy of model averaging. The process is briefly outlined in Section 4.2, under the subsection relative to results with neural networks. Prior to the training phase, all predictors were centered and scaled.

The weight decay (parameter *decay*) was tested from a list with the values 0.001, 0.01 and 0.1, while the number of hidden units (parameter size) was varied from a sequence initiated at 1 and increased two units until 27. Resampling results for the generated models are presented in Table 5.6. Figure 5.4 illustrates how the RMSE behaves in function of the parameters.

Table 5.6: Sample of the results for the neural network models for optional curricular units per semester

decay	size	RMSE	R^2
0.001	1	87.576	0.991
0.001	3	67.131	0.949
0.001	5	54.446	0.995
0.001	7	52.029	0.960
...	...	...	...
0.010	13	46.535	0.937
...	...	...	...
0.100	27	49.585	0.930

The final model, selected on the basis of RMSE, used 13 hidden units and had the weight decay set at 0.01.

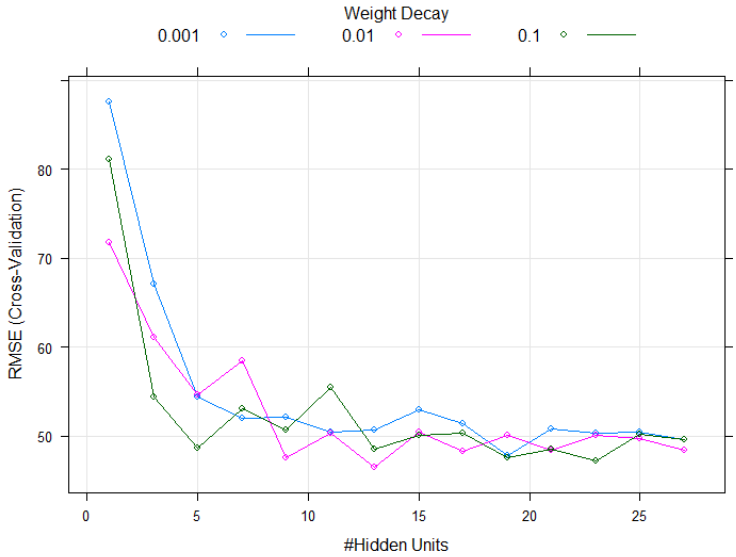

Figure 5.4: Neural network model variations for optional curricular units per semester

5.2.6 k-Nearest Neighbors

Regression models based on k-NN had the predictors centered and scaled prior to the training phase. Categorical variables were manually transformed using the dummies strategy. As with SVM, the removal of the *mobility* variables was found to produce a better fit. The number of neighbors averaged (parameter k) was varied between 1 and 15. Resampling results for the generated models are presented in Table 5.7. Figure 5.5 illustrates how the RMSE behaves in function of the parameters.

Table 5.7: Sample of the results for the KNN models for optional curricular units per semester

k	RMSE	R^2
1	45.661	0.954
2	31.997	0.954
3	48.013	0.957
4	59.610	0.999
5	80.109	0.997
...	...	...
15	129.797	0.876

In all the instances with more than two neighbors observed, the average error increased as more neighbors were added. RMSE was used to select the optimal model, with the number of neighbors set at 2.

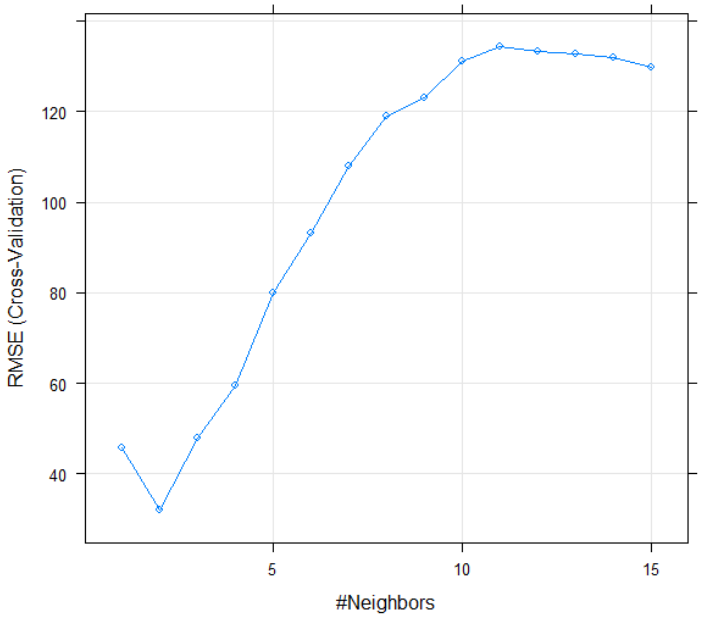

Figure 5.5: KNN model variations for optional curricular units per semester

5.2.7 Model and Rules Trees

The modeling approach selected for this topic is based on the M5 algorithm, with no predictor transformations applied. All combinations of pruning and smoothing processes were tested. Resampling results for the generated models are presented in Table 5.8. Figure 5.6 illustrates how the RMSE behaves in function of the parameters.

Table 5.8: Results for the M5 models for optional curricular units per semester

prune	smoothed	rules	RMSE	R^2
Yes	Yes	Yes	82.836	0.989
Yes	Yes	No	76.454	0.989
Yes	No	Yes	83.973	0.998
Yes	No	No	92.383	0.997
No	Yes	Yes	93.844	0.908
No	Yes	No	75.307	0.989
No	No	Yes	82.969	0.969
No	No	No	97.120	0.999

The final model, selected on the basis of RMSE, used a smoothing process and no pruning.

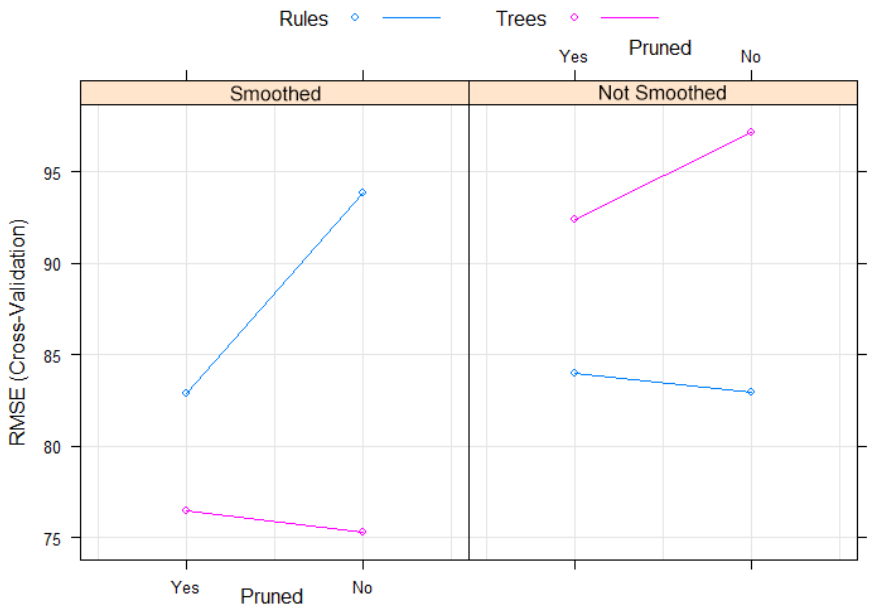

Figure 5.6: M5 model variations for optional curricular units per semester

5.2.8 Cubist

The process conducted for the generation of the Cubist models here presented did not require any variable transformations. The number of committees and neighbors, represented by parameters of identical names, were taken from lists with the values 1, 5, 10, 50, 75 and 0, 1, 3, 5, 7, 9, respectively. Resampling results for the generated models are presented in Table 5.9. Figure 5.7 illustrates how the RMSE behaves in function of the parameters.

Table 5.9: Sample of the results for the cubist models for optional curricular units per semester

committees	neighbors	RMSE	R^2
1	0	87.618	0.997
1	1	89.218	1.000
1	3	86.309	0.996
1	5	87.914	0.998
...	...	...	...
50	3	63.838	0.9986
...	...	...	...
75	9	67.165	0.999

The final model, selected using the RMSE as the evaluation metric, utilized 50 committees and 3 neighbors.

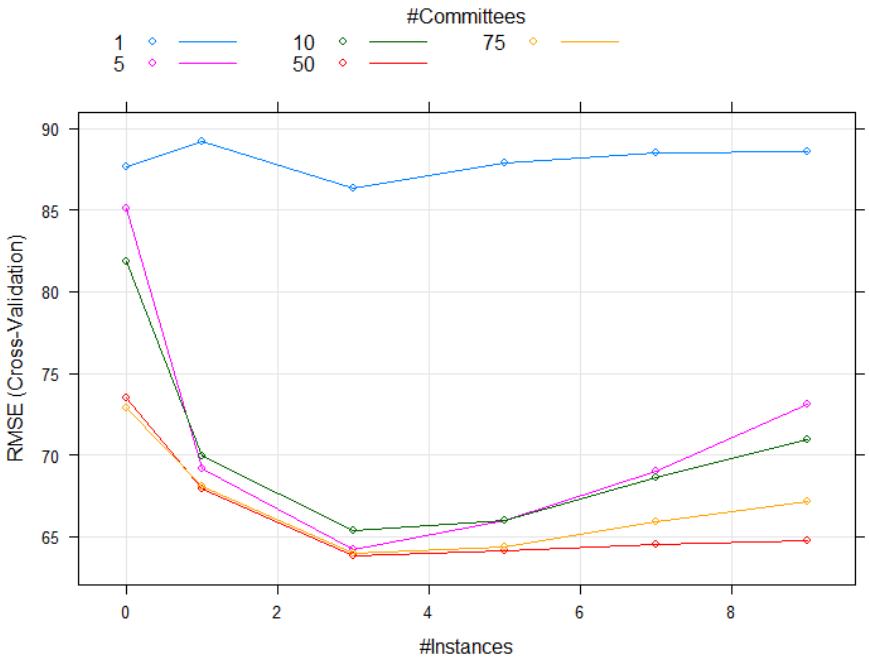

Figure 5.7: Cubist model variations for optional curricular units per semester

5.2.9 Random Forests

Regression models based on random forests were generated with no predictor transformations. The number of randomly selected predictors at each split (parameter *mtry*) was varied between 2 and the total number of predictors contained in each sample. Resampling results for the generated models are presented in Table 5.10. Figure 5.8 illustrates how the RMSE behaves in function of the parameters.

Table 5.10: Results for the random forests models for optional curricular units per semester

mtry	RMSE	R^2
2	91.632	0.998
3	85.094	0.985
4	82.234	0.986
5	82.110	0.988
6	79.917	0.983

Note that the results here presented refer to a forest with 500 trees. Additional tests with 1000 trees and increments of 1000 up to a total of 5000 trees yielded no improvements. RMSE was used to select the optimal model, with *mtry* set at 6 predictors.

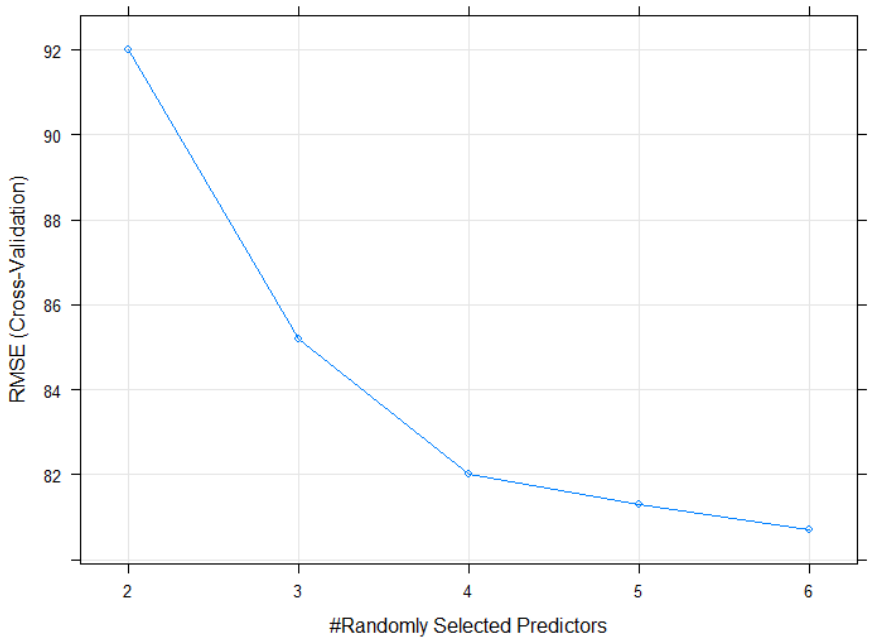

Figure 5.8: Random forest model variations for optional curricular units per semester

5.2.10 Aggregated Results

Overall, the models presented in this section managed to provide a decent estimate of the number of students enrolling in optional curricular units. Table 5.11 presents a brief summary, ordered by RMSE, of the resampling results obtained with 10-fold cross-validation for the best models constructed for each regression algorithm. k-NN and MARS were found to obtain the best results in regards to average error, with MARS achieving a higher proportion of variance explained. Linear regression models placed similarly, indicating that the prediction topic can be accurately described by a linear combination of the predictor variables. However, model trees constructed from multiple linear models placed much further.

Unlike the results presented in Chapter 4, the performance generally differs significantly between the results in each half of the models. As previously explained, this may be a consequence of the relatively small number of observations in the dataset, as some predictive models are naturally better suited to handle fewer samples.

Table 5.11: Summary of the results for all models tested for optional curricular units per semester

	RMSE	R^2
k-NN	32.00	0.9543
MARS	33.25	0.9998
Linear Regression	41.86	0.9795
Partial Least Squares	44.23	0.9433
Neural Networks	46.53	0.9368
SVM	51.92	0.9998
Cubist	63.84	0.9986
Model Tree	75.31	0.9895
Random Forests	79.92	0.9828

An additional test, similar to the one presented in Section 4.2, was also conducted. The test scenario partitioned the data in two blocks: a training set, composed of the academic years of 2009/2010 to 2014/2015, and a test set pertaining to the academic year of 2015/2016. The train and test sets were built from 18 and 3 samples, respectively. The three best regression models previously obtained were compared between themselves and against other estimates.

The results for the test case are presented in Table 5.12. As depicted, all models managed to surpass the naive average baseline, proving their relative predictive success. However, no model displayed noticeable improvement from an average using only previous occurrences. k-NN was the only model in which the results differ distinctively from the resampled RMSE.

Table 5.12: Comparison between the three best regression models constructed for optional curricular units per semester and other estimates

Prediction	RMSE
k-NN	104.951
MARS	34.772
Linear Regression	28.017
Naive Average	125.838
Naive Previous	28.400

Nonetheless, it is essential to mention that, while the regression models are expected to maintain the average evidenced by the resampling results, the same does not apply to the estimates described. The naive approach based on previous occurrences, for instance, displays a RMSE of 63.515 and 29.611 when applied to the academic years of 2014/2015 and 2013/2014, respectively.

Interestingly, both MARS and the linear regression models identify the three most significant predictors as *curricular.year*, *semester* and *students*. The variables pertaining to *mobility* have comparatively low coefficients in the linear regression model, and are not utilized in MARS. The fact that both models recognize the importance of the curricular year and semester also serves to demonstrate the reason as to why a simple average using only previous occurrences (from the same curricular year and semester) achieves comparatively positive results.

5.3 Experiments

After concluding the modeling phase, two additional experiments were developed so as to ascertain the possible advantages of introducing new predictors that reflect information from previous occurrences. The only models considered in the experiment were k-NN, MARS and ordinary linear regression, the ones which attained the lower average error in the previous step. All the results examined in this section were derived after a process of 10-fold cross-validation.

5.3.1 Prediction from Previous Occurrences

As evidenced, the significance of previous occurrences of a given observation is such that a prediction based on a naive average is capable of achieving results competitive with those obtained by a complete predictive model. Therefore, it is expected that the inclusion of predictors describing previous occurrences may prove capable of improving the regression models.

The format analyzed is heavily influenced by the structure utilized for the prediction topic inspected in Chapter 4. Four new independent variables were introduced: *prev.optionals.registered*, *mean.optionals.registered*, *prev.students* and *students.change*. Entries preceded by *prev* indicate information pertaining to an observation's previous occurrence. The variable *mean.optionals.registered*

is given by the mean number of students registered in optional curricular units in a sample's previous occurrences. The predictor *students.change* was calculated as the ratio between the number of students registered in the sample's semester and the number of students registered in the previous occurrence. The predictors pertaining to *mobility* were removed due to their reduced impact in the results.

The first observations in the dataset had no information about previous occurrences and were, thus, removed from the experiment. The final dataset was comprised of 18 samples. Table 5.13 presents a comparison of the results obtained with the three best models utilizing both formats. For all cases, the RMSE was calculated after a process of 10-fold cross-validation.

Table 5.13: Comparison between the three best regression models constructed for optional curricular units per semester utilizing extra predictors based on previous occurrences

Prediction	RMSE - Original Structure	RMSE - Experimental Structure
k-NN	32.00	33.92
MARS	33.25	32.50
Linear Regression	41.86	29.49

As depicted, the addition of new predictors significantly improves the linear regression model. An analysis of the corresponding variable significance concludes that, while the model maintains its three most important variables, the new predictors also contribute towards the response. The MARS model may be seen as an entirely new alternative, as it only utilizes the variable *mean.optionals.registered*. The k-NN model is not improved.

5.3.2 Exhaustive Search of Predictor Combinations

As an extension of the previous experiment, an exhaustive search was conducted so as to examine which predictors have the most impact in the prediction process. Table 5.14 presents a comparison of the results obtained with the best combinations of variables found for k-NN, MARS and ordinary linear regression.

Table 5.14: Comparison between the three best regression models constructed for optional curricular units per semester utilizing new predictor combinations

Prediction	RMSE - Original Structure	RMSE - Experimental Structure
k-NN	32.00	26.15
MARS	33.25	30.23
Linear Regression	41.86	21.74

Although the new k-NN and MARS models see little change, the new linear regression model displays remarkable improvement. The updated version drops four variables, being left with the

following: *year*, *curricular.year*, *semester* and *students.change*. In summary, when compared to the original structure, the only addition is *students.change*. Due to the removal of predictors related to *mobility*, the new model uses two less independent variables.

When evaluated under the test scenario's conditions, the updated linear regression model achieves a RMSE of 13.45. Compared to the naive prediction based on previous occurrences, this represents an error less than twice as small. Nonetheless, it should be noted that this is not the average outcome. When arguing about statistical significance, the results obtained with cross-validation should always take precedence.

5.4 Conclusions

Based on the resampling procedure selected for model evaluation and performance estimation, it can be inferred that the best regression method constructed for this topic is one based on **ordinary linear regression**. The model, presented in detail in Section 5.3, utilizes the predictors *year*, *curricular.year*, *semester* and *students.change*. After a resampling process of 10-fold cross-validation, the model is estimated to have a RMSE of 21.74, surpassing any naive prediction method.

When analyzing the error in the problem's context, this number represents roughly one class. While this result is far from optimal, the predictive model is expected to improve as more data is added and the variable's coefficients are arranged. At the moment, although the dataset contains data from six academic years, the information is only translated to 18 observations. In a machine learning problem, it not uncommon to have samples in the orders of hundreds or thousands of units.

Overall, however, given the problem's context and existing alternatives, the regression model here presented is shown to adequately infer new predictions by capturing internal relationships in the data. When considering the mean response of 205.8 registrations per semester, the proposed model achieves an average error of around 10%.

Chapter 6

Number of Students per Optional Curricular Unit

This chapter presents the results obtained with the predictive models constructed for the prediction of the number of student applications per optional curricular unit. Section 6.1 describes, in detail, the data structure selected for the topic. Section 6.2 illustrates the process applied to and results obtained by each individual model, and an additional experiment is discussed in Section 6.3. Conclusions on which model provides a better fit to the problem are presented in Section 6.4.

6.1 Experimental Setup

The original dataset constructed for curricular units included data relative to all units from the academic years of 2009/2010 to 2015/2016. However, information pertaining to student applications per optional unit was only available for the academic years of 2014/2015 and 2015/2016. The number of applications is the response variable in this prediction topic and must, thus, be present in every sample. As such, the dataset had to be pre-processed so as to only include data from 2014/2015 and 2015/2016. This filtering process shortened the number of observations from 210 to 54.

The initial format proposed for this topic was constructed on the basis that the previous occurrence of a given curricular unit has direct influence in the future edition. This assumes that factors such as average student grade and number of registrations has an impact in the subjects chosen by students during the selection process. The predictor and response variables are presented in Table 6.1, where the entries preceded by *prev* indicate information pertaining to the subject's previous occurrence.

The variable *prev.fails* was not directly available in the dataset, and was calculated as the number of students registered in the previous occurrence minus the ones who were approved. Similarly, the variable *registered.semester.change* was calculated as the ratio between students registered in the sample's semester and students registered in the previous occurrence. Both predictors, direct

correlations between original variables, were introduced due to their potential influence in the response variable.

Table 6.1: Data structure used in models for optional curricular units

Predictors			Response		
Variables	Mean	SD	Variable	Mean	SD
code	✗	✗			
year	✗	✗			
curricular.year	✗	✗			
semester	✗	✗			
prev.registered	19.9	7.9			
prev.evaluated	17.9	7.6			
prev.approved	17.5	7.8			
prev.fails	1.1	1.0	candidates	30.4	19.2
prev.evaluated.avg	15.3	1.7			
prev.evaluated.sd	1.6	0.8			
prev.approved.avg	15.3	1.6			
prev.approved.sd	1.6	0.7			
prev.registered.semester	527.7	112.3			
registered.semester	564.2	104.8			
registered.semester.change	1.1	0.3			
teacher	✗	✗			

The predictors *code*, *curricular.year*, *semester* and *teacher* were converted to factors; all other predictors were represented as numbers. The categorical variables contained 30, 2, 2 and 24 levels, respectively.

6.2 Results

The results, as presented by *caret*, are displayed over the following sections. All models were constructed under the same conditions in regards to environment and sampling partitions using a total of 54 samples and 16 predictor variables. The results here presented were obtained after a process of 10-fold cross-validation with each block estimated to have between 47 and 49 samples.

The selection of the models depicted in this topic was based on modeling interpretability. The regression models adopted were ordinary linear regression and CART. MARS was later elected due to its predictive capabilities.

6.2.1 Ordinary Linear Regression

Initially, when attempting to train the predictive model with all defined predictors, R returned multiple warnings with the message "prediction from a rank-deficient fit may be misleading". Possible causes for this issue are clarified in Section 4.2. In this case, the total number of levels in categorical variables surpassed the number of observations. In some cross-validation folds, predictions failed entirely. In R's implementation, attempting to predict samples with new levels for categorical predictors results in a failure, as those levels were not seen during the training phase and the model does not know how to adapt. Due to the reduced number of observations, some curricular units are only represented in the dataset by a single sample. When those samples are randomly selected to be part of the testing block, seeing as their code is unique, predictions cannot be computed. Note that, as the generation of the training and testing subsamples are conducted within *caret*, it is not possible to manually set the levels for the training blocks.

The preliminary model resulted in a RMSE of 25.261 and a R^2 of 0.352. Once the categorical variables *code*, *curricular.year* and *teacher* were removed from the model, the warnings disappeared and the RMSE was reduced to 14.683. Additional experiments were focused on studying variable importance so as to remove unnecessary predictors. This process was based on the absolute value of the *t*-statistic for each parameter. The final model utilized a total of 6 predictors, shortened from the original 16. Its results are presented in Table 6.2.

Table 6.2: Results for the linear regression model selected for optional curricular units

RMSE	R^2
13.379	0.581

The dependent variables selected for the model, along with the regression intercept, are displayed in Table 6.3. For each predictor, the table presents its value in the regression equation, known as the coefficient, and corresponding standard error. The rightmost column, obtained with a *t*-test, can be interpreted as the level of confidence in the hypothesis that the corresponding predictor is null. For instance, a value of 0.01 would represent a confidence value of 99% that the coefficient is not null. Variables with a symbol in front of them have the null hypothesis rejected with over 90% confidence.

Table 6.3: Predictors used in the linear regression model constructed for optional curricular units

Variable	Coefficient	Std. Error	$Pr(>\|t\|)$
Intercept	-11460.824	7332.821	0.12477
year	5.654	3.641	0.12719
semester2S	6.379	3.739	0.09458 .
prev.registered	-1.600	1.001	0.11673
prev.evaluated	2.584	1.092	0.02217 *
prev.evaluated.avg	4.561	1.186	0.00036 ***
registered.semester.change	12.890	6.533	0.05438 .
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Note that the removal of further variables lead to an increase in the fitting error of the model.

6.2.2 Basic Regression Trees

The complexity parameter utilized in the pruning phase (parameter *cp*) was varied from a sequence of thirty equally separated values between 0.00 and 0.57. Resampling results for the generated models are presented in Table 6.4.

Table 6.4: Sample of the results for the CART models for optional curricular units

cp	RMSE	R^2
0.000	15.920	0.365
0.020	15.922	0.366
0.040	16.156	0.406
0.060	16.918	0.370
0.080	16.778	0.379
..	...	...
0.578	19.612	0.258

As evidenced, the increase of the complexity parameter resulted in a higher average error. RMSE was used to select the optimal model, with the complexity set at 0.0. The resulting tree exhibited a total of 9 nodes, with 5 leaves. It is illustrated in Figure 6.1.

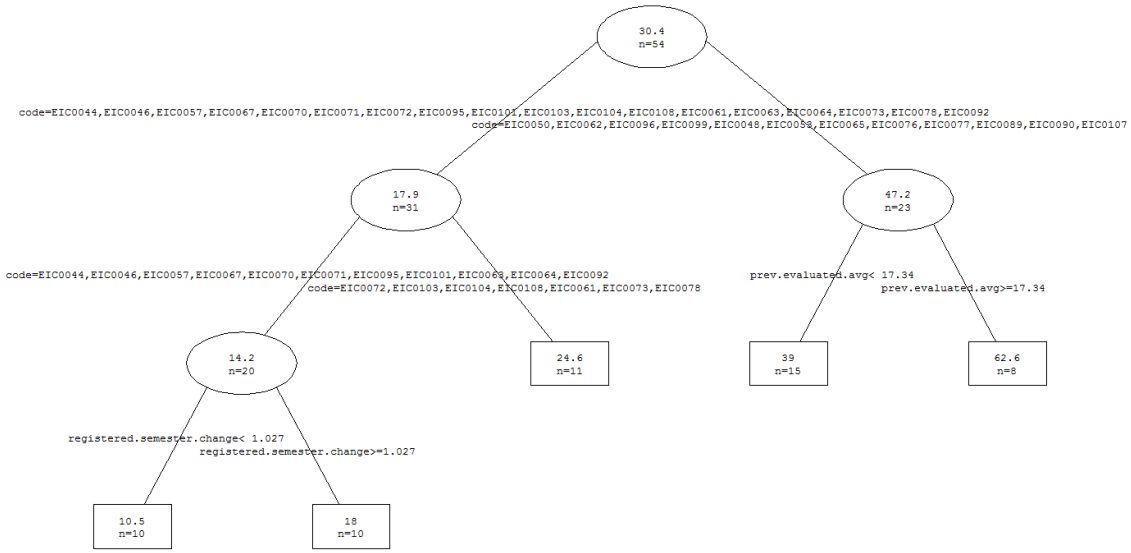

Figure 6.1: Tree model for optional curricular units

The variables selected for the splits were *code*, *registered.semester.change* and *prev.evaluated.avg*.

6.2.3 Multivariate Adaptive Regression Splines

The number of terms maintained (parameter *nprune*) was varied between 2 and 25, while the product degree (parameter *degree*) was held constant at a value of 1. Resampling results for the generated models are presented in Table 6.5. Figure 6.2 illustrates how the RMSE behaves in function of the parameters.

Table 6.5: Sample of the results for the MARS models for optional curricular units

nprune	RMSE	R^2
2	14.001	0.525
3	13.609	0.553
4	14.043	0.509
5	14.646	0.449
6	14.303	0.454
...	...	...
25	15.123	0.453

The average error was found to increase as more terms were maintained, stabilizing after 13 terms. The final model, selected on the basis of RMSE, had *nprune* set at 3.

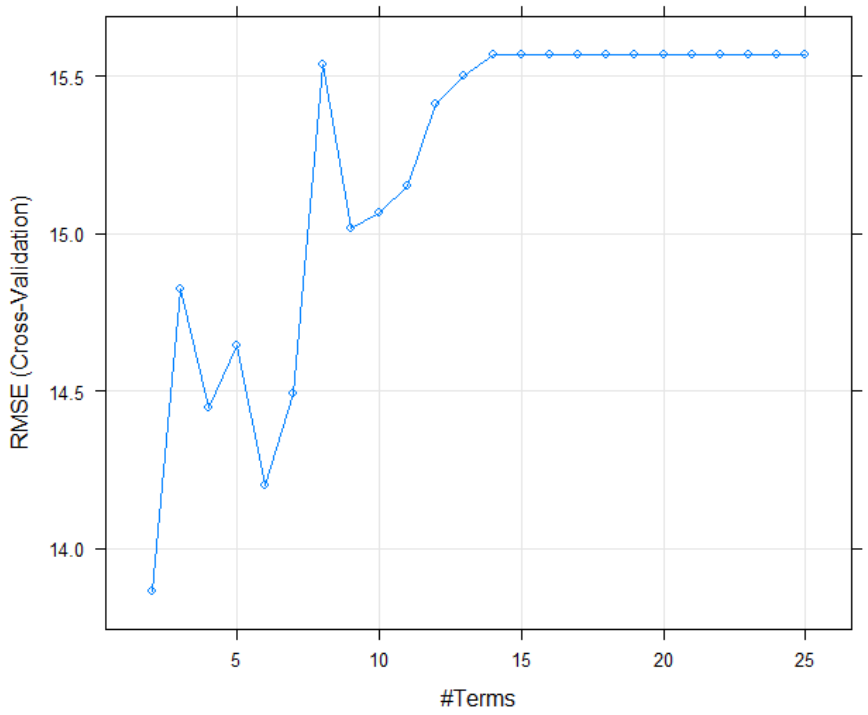

Figure 6.2: MARS model variations for optional curricular units

The predictors selected by the final model were *year* and *prev.evaluated.avg*. As a binary variable with only two possible values (2014 and 2015), *year* did not utilize a cut point; the cut point for *prev.evaluated.avg* was calculated to be at 16.61.

6.2.4 Aggregated Results

Overall, when considering the mean response of 30.4 candidates per optional curricular unit, the models failed to provide an estimate of the number of students enrolling in optional curricular units with an average error under 30%. Table 6.6 presents a brief summary, ordered by RMSE, of the resampling results obtained with 10-fold cross-validation for the best models constructed for each regression algorithm. The linear regression model was found to obtain the best results in regards to both average error and proportion of variance explained, with MARS falling slightly behind.

Table 6.6: Summary of the results for all models tested for optional curricular units

	RMSE	R^2
Linear Regression	13.379	0.581
MARS	13.609	0.553
CART	15.920	0.365

An additional test, similar to the ones presented in Sections 4.2 and 5.2, was also conducted. The test scenario partitioned the data in two blocks: a training set, composed of the academic year of 2014/2015, and a test set pertaining to the academic year of 2015/2016. The train and test sets were built from 30 and 24 samples, respectively.

Results for the test case are presented in Table 6.7. As depicted, all three models managed to surpass the naive baseline, proving their relative predictive success. However, no model displayed noticeable improvement from the estimates. Possible causes for this fact are identified and explained in Section 6.4.

Table 6.7: Comparison between the regression models constructed for optional curricular units and other estimates

Prediction	RMSE
Linear Regression	16.146
MARS	18.907
CART	18.484
Naive Average	19.766
Naive Previous	18.838

Interestingly, despite the relatively lacking results, all the regression models analyzed determined the average grade of the evaluated students in the unit's previous occurrence to be a significant factor in the predictions. Both the ordinary linear regression and the CART models also recognized the difference in the number of students per semester to be a relevant variable.

6.3 Experiments

After concluding the modeling phase, an additional experiment was developed so as to ascertain the possible advantages of introducing new predictors that reflect information from student questionnaires. All the results examined in this section were derived after a process of 10-fold cross-validation.

6.3.1 Prediction From Student Questionnaires

Twice a year, at the end of a semester, FEUP asks its students to fill in questionnaires regarding the curricular units they were registered in. The questionnaires include a multitude of topics, from the subject's difficulty to the instructor's relationship with the students. The topics approached in this experiment were related to the unit itself. They were as follows:

- Appreciation and clarity;
- Evaluation;
- Difficulty;
- Impact.

The second topic relates to how a student perceives the unit's grading policy in regards to, for instance, the number of tests and their influence in the final grade. Impact illustrates the influence of the subject in a student. The topics of appreciation and difficulty are self-explanatory.

All topics were rated by students in a scale of 1 to 7 points. It is important to note that the questionnaires are optional and, for the majority of the curricular units, completed by less than 30% of the student body. This information was not included as part of the predictor variables.

For a given entry in the dataset, each questionnaire topic was set as the average response to the topic related to the unit's previous occurrence. As an example, the unit X in the academic year of 2014/2015 would have its appreciation topic set as the the average response given in the questionnaire related to the same unit in 2013/2014. The corresponding predictor variables were introduced as *prev.questionnaire.appr*, *prev.questionnaire.eval*, *prev.questionnaire.diff* and *prev.questionnaire.imp*, following the structure presented in Section 6.1.

Table 6.8: Comparison between the regression models constructed for optional curricular units utilizing extra predictors for student questionnaires

Prediction	RMSE - Original Structure	RMSE - Experimental Structure
Linear Regression	13.379	12.403
MARS	13.609	13.375
CART	15.920	15.920

Table 6.8 presents a comparison of the results obtained with the experimental format. The CART model displayed no difference when trained with the questionnaire variables. MARS maintained new terms related to *prev.questionnaire.appr* and *prev.questionnaire.diff*, but their inclusion did not improve the fit in a substantial way. The ordinary linear regression model, however, showed significant improvement.

The linear regression model utilized the predictors *prev.questionnaire.appr*, *prev.questionnaire.eval* and *prev.questionnaire.diff*, as the removal of the variable *prev.questionnaire.imp* was shown to be

beneficial. A new t-test demonstrated that, for all questionnaire topics maintained, the null hypothesis could be rejected with over 97% confidence. The new model achieved a R^2 value of 0.687, improving the initial value of 0.581. When evaluated under the test scenario's conditions, the RMSE obtained was 18.559.

6.4 Conclusions

Based on the resampling procedure selected for model evaluation and performance estimation, it can be inferred that the best regression method constructed for this topic is one based on **ordinary linear regression**. The model, presented in detail in Section 6.3, utilizes a total of 9 predictors. After a resampling process of 10-fold cross-validation, the model is estimated to have a RMSE of 12.403, surpassing any naive estimate.

All models suggest the average grade of the students in the unit's previous occurrence as a significant factor in the predictions. It is dangerous to interpret this a direct causation, however, as it is possible that students are simply more likely to select and perform well in subjects they prefer. The number of students registered in the semester of the the unit's occurrence is also seen as a potential influence. This may be explained by the fact that more students in a semester implies an increase in the overall number of applications.

Despite the relatively positive results, the models analyzed are not yet capable of reliably estimating the number of applications per optional curricular unit. This paper proposes three possible causes for these results, detailed below:

- **Insufficient data**: the number of observations was directly responsible for several modeling decisions that took place during the implementation phase. Categorical variables such as a unit's code, removed from the linear regression model, could greatly improve the overall fit. This hypothesis is also supported by the CART model, which uses the code as the initial split. Likewise, the influence of variables such as the instructor cannot be accurately calculated when they are only represented in one sample of the dataset.

- **Lack of relevant predictors**: it is possible that the predictors under analysis fail to entirely capture the factors underlying a student's choice. It could be advantageous to identify other distinguishing traits among units, such as area of study or grading policy (with or without exam, for instance).

- **Unpredictable relations**: the selection process may be influenced by factors outside the unit's scope, such as a student's characteristics. Attempting to predict which subjects a student will select based on the student's age or academic average, rather than attempting predictions on a macro-scale, could prove beneficial.

Overall, despite their shortcomings, the regression models represent the best predictive alternative analyzed in this study. It is important to clarify that this study's objective is both predictive performance and variable interpretability, so as to assist a course director in the understanding

of student demands. In general, the models presented prove their usefulness as tools capable of providing the course's administration with data that can potentially support informed decisions.

Chapter 7

Conclusions

Every year, all around the world, millions of students are required to choose the curricular units they are interested in enrolling for the coming semesters. In order to successfully plan a scholar year, higher education institutes aim to accurately predict and understand their students' demands. This dissertation presented an approach based on predictive analytics on how to support the administrative necessities of a course director. Three predictions topics were analyzed: number of students per non-optional curricular unit, number of students enrolling in optional curricular units, and number of students per optional curricular unit. Topics were examined separately, and different predictive models were formulated for each case. To validate the hypothesis here presented, the models developed were applied, in the form of a case study, to the MIEIC course at FEUP.

On a macro perspective, this study demonstrated that the application of data mining methodologies surpassed any estimate currently used by the course's administration. For each topic, the final regression model presented was shown to perform better than naive estimates calculated from previous occurrences of curricular units or semesters and their averages. In order to reliably estimate the performance of the predictive models, results were validated using k-fold cross-validation.

For the topic of students per non-optional curricular unit, each sample was structured so as to include statistics from the unit's previous occurrence due to the influence of data sequences. Given a mean response of 152.6, MARS proved to achieve the best results, presenting a RMSE value of 9.006. Further experiments using ensemble methods managed to lower the average error to 8.868, using a GLM composed of a MARS, cubist and boosting models.

On the prediction of students enrolling in optional curricular units, given a mean response of 205.8, an ordinary linear regression model was estimated to have a RMSE value of 21.74. The features pertaining to the number of students in mobility programmes initially contemplated were removed from the final model due to their reduced contribution to the results.

Lastly, for the topic of student applications per optional curricular unit, given a mean response of 30.4, an ordinary linear regression model was shown to achieve a RMSE value of 12.403.

Proportionally, although this represents a much higher average error than on the other prediction areas, it illustrates an average error reduction of around 35% when compared to naive baselines estimated using the outcomes of previous occurrences. When reviewing the agglomeration of models constructed, the average grade of the students evaluated in the unit's previous occurrence and number of students per semester are suggested as significant factors in the predictions.

7.1 Future Work

The application of predictive analytics to the modeling of student populations, in particular when focused on the scope of individual curricular units, is a topic that has yet to be thoroughly explored. The area of Educational Data Mining, in which this study was conceived, is by itself an emerging discipline that has just began growing over the last few years. There are various aspects that could be considered in future follow-ups, to note:

- Improvement of the original dataset. The data provided for the case study was fairly fragmented, with multiple features holding values from distinct time ranges. Increasing the total number of observations so as to reflect all occurrences from the academic years of 2009/2010 to 2015/16 could improve model reliability.
- Further experimentation with student questionnaires. As evidenced in Chapter 6, questionnaires may be used as predictor variables with predictive potential. Additional experiments could include extra questionnaire topics and the number of responses associated with each topic. This could allow algorithms to adjust to cases where the percentage of questionnaire responses is not statistically relevant.
- Extension of the case study to other courses and faculties. At the moment, it is not possible to estimate the predictive performance of the models constructed on other courses. Further experiences could demonstrate how individual features and models generalize and adapt to other circumstances.
- Prediction of the number of student registrations, and not applications, per optional curricular unit. Assuming a maximum of one class with a limited number of students per curricular unit, the number of applications may prove irrelevant. For instance, given a maximum of 20 students per class, two units with 20 and 50 applications, respectively, will both have the same number of registrations. As such, this might prove a viable alternative on how to predict which units to allocate per semester.
- Predictions on a student by student basis. Attempting to estimate which subjects a given student will enroll in might also produce viable predictions. While this methodology has already been applied to the topic of student attrition, it has yet to be applied within this scope. Rather than a direct continuation of the case study, this topic should be treated as an extension to the investigation of predictive analytics on modeling student populations.

References

[BY09] Ryan S J D Baker and Kalina Yacef. The state of educational data mining in 2009 : A review and future visions. *Journal of Educational Data Mining*, 1(1):3–16, 2009.

[CD14] T. Chai and R. R. Draxler. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, jun 2014.

[CDO07] J P Campbell, P B DeBlois, and D G Oblinger. Academic analytics. *Educause Review*, 42(October):40–57, 2007.

[Del11] Dursun Delen. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, 13(1):17–35, 11 2011.

[DMK16] Zachary A. Deane-Mayer and Jared E. Knowles. caretEnsemble. 2016.

[DPV09] Gerben W. Dekker, Mykola Pechenizkiy, and Jan M. Vleeshouwers. Predicting students drop out: A case study. *EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining*, pages 41–50, 2009.

[FPSS96] Usama Fayyad, G Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, pages 37–54, 1996.

[HHWZ08] Torsten Hothorn, Kurt Hornik, Wirtschaftsuniversitat Wien, and Achim Zeileis. Party: A Laboratory for Recursive Partytioning. jan 2008.

[HHZ] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. ctree : Conditional Inference Trees.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. *Elements*, 1:337–387, 2009.

[Hue13] Richard Huebner. A survey of educational data-mining research. *Nature*, 194(4833):1006–1006, 2013.

[IR07] Zaidah Ibrahim and Daliela Rusli. Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression. *Proceedings of the 21st Annual SAS Malaysia Forum*, (September):1–6, 2007.

[Jin02] Luan Jing. Data mining and knowledge management in higher education. *Workshop associate of institutional research international conference, Toronto*, pages 1–18, 2002.

[KJ13] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. 2013.

[KMZ00] Marie Khair, Chady El Moucary, and Walid Zakhem. Creating an educational roadmap for engineering students via an optimal and iterative yearly regression tree using data mining. pages 43–52, 2000.

[KSGS13] Ahmad A. Kardan, Hamid Sadeghi, Saeed Shiry Ghidary, and Mohammad Reza Fani Sani. Prediction of student course selection in online higher education institutes using neural network. *Computers and Education*, 65:1–11, 2013.

[KWKC12] Max Kuhn, Steve Weston, Chris Keefer, and Nathan Coulter. Cubist Models For Regression, 2012.

[Lab15] Cybermetrics Lab. Countries arranged by number of universities in top ranks, 2015.

[LG08] Gerard Lassibille and Lucia Navarro Gomez. Why do higher education students drop out? evidence from spain. *Education Economics*, 16(1):89–105, 2 2008.

[LLH12] Jin-Ling Lin, Jy-Hsin Lin, and Kao-Shing Hwang. The number of students needed for undecided programs at a college from the supply-chain viewpoint. *Mathematical Problems in Engineering*, 2012, 2012.

[Lua02] Jing Luan. Data mining and its applications in higher education. *New Directions for Institutional Research*, 2002(113):17–36, 21 2002.

[LW14] Johann Ari Larusson and Brandon White. *Learning Analytics: From Research to Practice*. Springer, 2014.

[MBS99] P a Murtaugh, L D Burns, and J Schuster. Predicting the retention of university students. *Research in Higher Education*, 40(3):355–371, 1999.

[MMSJS12] João Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble Approaches for Regression: A Survey. *ACM Computing Surveys*, 45(1):1–40, 2012.

[MWK+16] Kuhn Max, Steve Weston, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, and Can Candan. caret. 2016.

[PP13] Nichaphat Patanarapeelert and Klot Patanarapeelert. Forecasting number of students in university department: Modeling approach. *Open Journal of Applied Sciences*, 03(04):293–297, 2013.

[RP13] Cecilia Rosa and Edgar Pereira. A metapopulation model for the study of the evolution of the number of students in a teaching institution. *Iberian Conference on Information Systems and Technologies, CISTI*, 2013.

[RV07] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007.

[RV10] Cristbal Romero and Sebastin Ventura. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 40(6):601–618, 2010.

REFERENCES

[SB12] George Siemens and Ryan S. J. d. Baker. Learning analytics and educational data mining: Torwards communication and collaboration. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*, pages 252–254, 2012.

[Sou15] Harry Southworth. gbm: Generalized Boosted Regression Models. page 34, 2015.

[Sta10] Jon Starkweather. Categorical Variables in Regression: Implementation and Interpretation. *University of North Texas*, 2, 2010.

[Str16] Pedro Strecht. Merging decision trees : an application to student performance. (January), 2016.

[Tin75] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, 1975.

[Tor10] Luis Torgo. *Data Mining with R: Learning with Case Studies*. 2010.

[VS11] W Venables and D Smith. An Introduction to R. 0, 2011.

[WF05] Ian H. Witten and Eibe Frank. Data mining: Practical machine learning tools and techniques (second edition). 6 2005.